Counter Free Regular Languages Are Not Learnable

Marcus Kracht
Department of Linguistics, UCLA
3125 Campbell Hall
Los Angeles, CA 90095-1543
kracht@humnet.ucla.edu

June 10, 2007

Dedicated to András Kornai on occasion of his 50th birthday

A language $L \subseteq A^*$ is **counter free** if there is a $k \in \mathbb{N}$ such that for all $\vec{x} \in L$, if $\vec{x} = \vec{u}\vec{v}^k\vec{w}$ then $\vec{u}\vec{v}^{k+n}\vec{w} \in L$ for every $n \in \mathbb{N}$. We call k the **threshold** of L. A short reflection will show that it is enough to require that for all $\vec{u}$, $\vec{v}$ and $\vec{w}$: $\vec{u}\vec{v}^k\vec{w} \in L$ iff $\vec{u}\vec{v}^{k+1}\vec{w} \in L$. We shall denote the set of counter free languages over A of threshold k by $\mathbb{Q}_k(A)$. (The case $k = 0$ is trivial: $\mathbb{Q}_0(A) = \{A^*\}$.) Evidently, $\mathbb{Q}_k(A) \subsetneq \mathbb{Q}_{k+1}(A)$ for every k. That the inclusion is proper is easy to see; $\{\mathsf{a}^n : n < k+1\}$ is in $\mathbb{RQ}_{k+1}(A)$ but not in $\mathbb{RQ}_k(A)$. The class of regular languages in $\mathbb{Q}_k(A)$ is denoted by $\mathbb{RQ}_k(A)$. Also $\mathbb{RQ}_k(A) \subsetneq \mathbb{RQ}_{k+1}(A)$.

Not every counter free language is regular (examples will be shown below). There are only a few exceptions: if $|A| = 1$, say $A = \{\mathsf{a}\}$, then every counter free language is either finite or cofinite, and hence regular. Given a threshold k, $\mathbb{RQ}_k(A)$ consists of all unions of the languages $\{\mathsf{a}^i\}$ $(i < k)$, $\{\mathsf{a}^{k+n} : n \in \mathbb{N}\}$, which are all regular. This class is learnable. Thus from now on we shall assume $|A| > 1$. Regular counter free languages can be characterised as the languages whose regular expression can be built from letters using union, concatenation and complement (normally, complement is not used since it is definable in presence of the star).

It has been claimed by András Kornai that natural languages are regular and counter free of threshold 4. It is thus of great interest to know whether such languages are learnable in the limit. Recall that a class $\mathcal{P}$ is learnable if there is a learning algorithm learning every member of $\mathcal{P}$. A **learning algorithm** is a computable function $\alpha : (\mathcal{P} \cup \{\varnothing\}) \times A^* \to \mathcal{P}$. (It is not assumed that $\varnothing \notin \mathcal{P}$.) For an infinite infinite sequence $\sigma = \langle \sigma_i : i \in \mathbb{N} \rangle$ define

$$L_0^\sigma := \varnothing, \qquad L_{i+1}^\sigma := \alpha(L_i^\sigma, \sigma_i) \tag{1}$$

α **learns** L **from** σ if L_i^σ is eventually constant and equals L. α **learns** L if it learns L from every sequence that contains every word of L at least once. Thus, learnability is a property of the class, not the individual languages, since α must be given for $\mathcal{P}$ as a whole. (There is a trivial algorithm that learns a single language.)

In this note we shall show that the regular counter free languages of threshold k are not learnable if the class is infinite. The proof is a direct application of results by Angluin on the existence of so-called telltales ([1]).

Definition 1 *Let $\mathcal{P}$ be a class of languages. A **telltale** for L in $\mathcal{P}$ is a finite set $D \subseteq L$ such that for all M in $\mathcal{P}$: if $D \subseteq M$ then $M \not\subset L$.*

If a computable telltale exists for all languages in $\mathcal{P}$ then all languages in $\mathcal{P}$ are identifiable in the limit.

Theorem 2 (Angluin) *If $\mathcal{P}$ is learnable then every language of $\mathcal{P}$ has a telltale.*

Proof. By assumption, the algorithm is eventually constant for given L and for all σ. Pick some σ and assume that L_p^σ is constant. Then $L_p^\sigma = L$. Moreover, $\alpha(L, \sigma_p) = L$. There is an alternative sequence σ' such that $\sigma'_i = \sigma_i$ for all $i < p$, and $\sigma'_p \in L$. Then we shall have $L_p^{\sigma'} = L$, so that $\alpha(L, \sigma'_p) = L$. This establishes that $\alpha(L, \vec{w}) = L$ for every $\vec{w} \in L$. Let $D = \{\sigma_i : i < p\}$. We show that D is a telltale for L. For suppose to the contrary that there is a $L' \in \mathcal{P}$ such that $D \subseteq L' \subset L$. Then there is a text τ for L' such that $\tau_i = \sigma_i$ for all $i < p$. Thus $L_p'^\tau = L$. But from what we have just seen, for every $\vec{w} \in L$ (and hence for every $\vec{w} \in L'$) $\alpha(L, \vec{w}) = L$, so that the algorithm is constant from this point on but returns the wrong language. □

The negative result now follows by establishing that there are languages in $\mathbb{RQ}_k(A)$ with $k > 1$ that have no telltale. Call a string $\vec{v}$ **square free** if it does not have the form $\vec{u}\vec{x}\vec{x}\vec{w}$ for any strings $\vec{u}$, $\vec{x}$ and $\vec{w}$. By a result of Axel Thue's (see [2]) there exists an infinite word $\xi = \langle x_i : i \in \mathbb{N}\rangle$ over an alphabet A with at least three letters which is square free. For a two letter alphabet, this is false. The only square free strings are ϵ, a, b, ab, ba, aba and bab. This is because repeated adjacent occurrences of the same letter are excluded so that the string must consist of a followed by b and conversely. If the string has length at least four, it contains either the sequence abab or the sequence baba. Thus in the case $|A| = 2$ and it turns out that $\mathbb{RQ}_2(A)$ (and thus also $\mathbb{RQ}_1(A)$) is learnable.

Now let ξ be given. We define $\xi_p := x_0 x_1 \cdots x_{p-1}$. The words ξ_p are also square free. We can now show that there are uncountably many counter free languages (and thus plenty of nonregular ones, as the regular ones are countable). Just take a subset $Q \subseteq \mathbb{N}$ and put $U_Q := \{\xi_q : q \in Q\}$.

Let L_i, $i \in I$, be languages from $\mathbb{Q}_k(A)$. Then the intersection $L := \bigcap_i L_i$ is also in $\mathbb{Q}_k(A)$. For suppose that $\vec{u}\vec{v}^*\vec{w} \in L$; then it is in L_i for any $i \in I$. Hence $\vec{u}\vec{v}^{k+1}\vec{w} \in L_i$. Since i was arbitrary, $\vec{u}\vec{v}^{k+1}\vec{w} \in L$. Hence we conclude that the following is well defined.

Definition 3 *Let H be a set of words. We denote by $[H]_k$ the smallest language in $\mathbb{Q}_k(A)$ containing H.*

We note that $\mathbb{RQ}_k(A)$ is closed under finite intersection only, so the argument above does not go through. It seems that $[H]_k$ is regular whenever H is finite, but we do not know how to prove that. $[H]_k$ can be created through an infinite process as follows. Let $H_0 := H$ and $H_{n+1} := (H_n)^+$, where

$$P^+ := P \cup \{\vec{u}\vec{v}^{k+1}\vec{w} : \vec{u}\vec{v}^*\vec{w} \in P\} \cup \{\vec{u}\vec{v}^*\vec{w} : \vec{u}\vec{v}^{k+1}\vec{w} \in P\} \tag{2}$$

Then

$$[H]_k := \bigcup_{n\in\mathbb{N}} H_n \tag{3}$$

It is easy to see that if $H = \{\vec{u}_i : i \in I\}$ then $[H]_k = \bigcup\{[\vec{u}_i]_k : i \in I\}$.

Lemma 4 *Let H be a finite set of words. Put $p := \max\{|\vec{u}| : \vec{u} \in H\}$. If $k > 1$ then $\xi_q \notin [H]_k$ for any $q > p$.*

Proof. By induction over the construction. By choice of q, $\xi_q \notin H_0$. Now suppose that $\xi_q \notin H_n$ but H_{n+1}. Then there is a $\vec{y} \in H_n$ such that it has the form $\vec{y} = \vec{u}\vec{v}^*\vec{w}$ and $\xi_q = \vec{u}\vec{v}^{k+1}\vec{w}$, or $\vec{y} = \vec{u}\vec{v}^{k+1}\vec{w}$ and $\xi_q = \vec{u}\vec{v}^*\vec{w}$. But ξ_q is square free, so neither case can arise. Contradiction. □

The previous lemma can be generalised.

Lemma 5 *Suppose that $\xi_q \notin H$ and $k > 1$. Then $\xi_q \notin [H]_k$.*

Proof. Similarly. □

We look at languages of the form $A^* - \{\xi_q : q \in P\}$ for some finite P. These languages are clearly regular; they are also in $\mathbb{Q}_k(A)$, by Lemma 5. It follows that the language A^* cannot be learned. This is made precise as follows.

Lemma 6 *Suppose that $|A| > 2$ and $k > 1$. Then A^* has no telltale in $\mathbb{RQ}_k(A)$.*

Proof. Let H be a finite set. We show that it is not a telltale for A^*. Pick $q > \max\{|\vec{u}| : \vec{u} \in H\}$. Then, by Lemma 5, the set $K := A^* - \{\xi_q\}$ is in $\mathbb{RQ}_k(A)$, and contains H. But $K \not\subset A^*$. □

We can sharpen this further.

Lemma 7 *The language $\{\xi_p : p \in \mathbb{N}\}$ has no telltale in $\mathbb{RQ}_1(A)$.*

Theorem 8 *The counter free regular languages satisfying for any given threshold $k > 0$ over an alphabet of size at least 3 are not learnable in the limit.* □

The case $|A| = 2$ remains to be dealt with. If $|A| = 2$ then there is an infinite sequence ξ not containing any subword more than three times. In this case, the previous proof can be redone. The demonstration is this. Consider a translation $\nu : \mathtt{a} \mapsto \mathtt{aa}$, $\mathtt{b} \mapsto \mathtt{bb}$ and $\mathtt{c} \mapsto \mathtt{ab}$. If ξ is square free then $\nu(\xi) := \langle \nu(\xi_i) : i \in \mathbb{N}\rangle$ does not contain any subword more than three times. ($\nu(\mathtt{ac}) = \mathtt{aaab}$, so this cannot be improved.) The argument is roughly this. Call an occurrence $\vec{v}$ in $\vec{w}$ **even** if it is preceded by a word of even length; and call it odd otherwise. The crucial fact is that the word $\nu(\vec{w})$ does not contain an even occurrence of $\mathtt{ba}$. Suppose that $\vec{x}$ is of even length, and suppose it has an even occurrence. Then $\vec{x} = \nu(\vec{y})$ for some $\vec{y}$. Now the word $\vec{w}$ is square free. And we have $\vec{x}\vec{x} = \nu(\vec{y}\vec{y})$, so if $\vec{x}$ is repeated in $\nu(\vec{w})$, $\vec{y}$ is repeated in $\vec{w}$. Now suppose that $\vec{x}$ has an odd occurrence. Suppose $\vec{x} = c\vec{u}d$. Put $\vec{y} := d\vec{x}c$. Then $\vec{x}\vec{x}\vec{x}$ contains an even occurrence of $\vec{y}\vec{y}$. This cannot be, as we have just seen. Now suppose that $\vec{x}$ is of odd length. And suppose that $\vec{x}\vec{x}$ is a subword of $\nu(\vec{w})$. If one of the occurrences contains $\vec{x}$ contains an odd occurrence of $\mathtt{ba}$, the other contains an even occurrence of $\mathtt{ba}$, which cannot

be. Hence $\vec{x}$ is of the form $\mathtt{a}^m\mathtt{b}^n$. In this case we can see that also $m < 4$ and $n < 4$. Suppose now that $m, n \neq 0$. Then $\vec{x}\vec{x}$ contains an occurrence of $\mathtt{ba}$, and $\vec{x}\vec{x}\vec{x}$ contains an even occurrence of $\mathtt{ba}$. Contradiction. So, $\vec{x} = \mathtt{a}^n$ or $\vec{x} = \mathtt{b}^n$ for $n < 4$. But then we have already seen that it cannot be repeated more than three times (and exactly three times for $n = 1$). The language $\{\nu(\xi_p) : p \in \mathbb{N}\}$ has no telltale in $\mathbb{RQ}_3(A)$. Since $\mathbb{RQ}_2(A)$ is finite, this exhausts all cases.

Theorem 9 *Let A be a finite alphabet and $k \in \mathbb{N}$, $k > 0$. Exactly one the following cases arises.*

① $|A| = 1$. *Then $\mathbb{RQ}_k(A)$ is finite.*

② $|A| = 2$, $k \leq 2$. *Then $\mathbb{RQ}_k(A)$ is finite.*

③ $|A| = 2$, $k > 2$. *Then $\mathbb{RQ}_k(A)$ is not learnable.*

④ $|A| > 2$. *Then $\mathbb{RQ}_k(A)$ is not learnable.*

References

[1] Dana Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135, 1980.

[2] Axel Thue. Über unendliche Zeichenreihen (on infinite strings). *Kra. Vidensk. Selsk. Skrifter I. Mat.-Nat. Kl.*, (7), 1906.