Matching Regular Pumping Lemmas and Automaticity

Karen Celine, NUS, Singapore

Ryan Chew, NUS, Singapore

Sanjay Jain, NUS, Singapore

Frank Stephan, NUS, Singapore

Talk for Logic Seminar 6 April 2022

Introduction

Traditional Regular Pumping Lemma [Michael Rabin and Dana Scott 1959, Yehoshua Bar-Hillel, Micha Perles and Eli Schamir 1961]
Given a regular language $\mathbf{L}$, there is a constant $\mathbf{k}$ such that for all words $\mathbf{u}$ of at least length $\mathbf{k}$, one can split $\mathbf{u}$ into $\mathbf{x}, \mathbf{y}, \mathbf{z}$ such that $\mathbf{y} \neq \varepsilon$, $|\mathbf{xy}| \leq \mathbf{k}$ for all $\mathbf{h} \in \mathbb{N}$, $\mathbf{L}(\mathbf{u}) = \mathbf{L}(\mathbf{xy^hz})$.

one-sided: Only $\mathbf{u} \in \mathbf{L}$ are considered;
two-sided: All $\mathbf{u}$ are considered.

The original formulation of the Pumping Lemma is one-sided.

Corollary (Weak Traditional Pumping Lemma)
One can in the above the condition $|\mathbf{xy}| \leq \mathbf{k}$ replace by the weaker condition $|\mathbf{y}| \leq \mathbf{k}$.

Example

Let $\mathbf{H}$ be a nonregular subset of $\{\mathbf{0},\mathbf{1},\mathbf{2}\}^*$ and $\mathbf{L} = \{\mathbf{0},\mathbf{1},\mathbf{22}\} \cdot \{\mathbf{0},\mathbf{1},\mathbf{2}\}^* \cup \{\mathbf{210}\} \cdot \mathbf{H}$. Now this language satisfies the two-sided pumping lemma but is not regular.

If a word $\mathbf{u}$ starts with one of $\mathbf{0},\mathbf{1},\mathbf{20},\mathbf{211},\mathbf{212},\mathbf{22}$ then pump the fourth symbol, which does not change the prefix and pumping lemma satisfied in this case.

If a word $\mathbf{u}$ starts with $\mathbf{210}$ and is in $\mathbf{L}$ then pump the first symbol, pumping down gives $\mathbf{1}$ as first symbol and pumping up gives $\mathbf{22}$ as starting symbols, so $\{\mathbf{2}\}^* \cdot \{\mathbf{10u}\} \subseteq \mathbf{L}$.

If a word $\mathbf{u}$ starts with $\mathbf{210}$ and is not in $\mathbf{L}$ then pump the second symbol, pumping down gives words starting with $\mathbf{20}$ which are not in $\mathbf{L}$ and pumping up gives words starting with $\mathbf{211}$. Thus $\{\mathbf{2}\} \cdot \{\mathbf{1}\}^* \cdot \{\mathbf{0u}\} \cap \mathbf{L} = \emptyset$.

$\mathbf{L} \cap \{\mathbf{210}\} \cdot \{\mathbf{0},\mathbf{1},\mathbf{2}\}^* = \{\mathbf{210}\} \cdot \mathbf{H}$ is nonregular, so is $\mathbf{L}$.

Matching Pumping Lemmas

Let $\mathbf{L_x} = \{\mathbf{y} : \mathbf{xy} \in \mathbf{L}\}$ be the derivative of $\mathbf{L}$ at $\mathbf{x}$. Myhill and Nerode investigated the derivatives and showed that a language is regular iff it has exactly finitely many distinct derivatives.

A pumping lemma is matching iff exactly the regular languages satisfy its pumping condition and no other languages do.

Theorem [Jaffe 1978].
The following two-sided pumping lemma is matching: There is a constant $\mathbf{k}$ such that all words $\mathbf{u}$ with $|\mathbf{u}| \geq \mathbf{k}$ can be split into $\mathbf{xyz}$ with $\mathbf{y} \neq \varepsilon$ and $\mathbf{L_{xyz}} = \mathbf{L_{xy^hz}}$ for all $\mathbf{h} \in \mathbb{N}$.

This is a two-sided pumping lemma and the one-sided version only pumping the words inside the language is not matching, example is $\mathbf{L} \cdot \{\mathbf{3}\}$ for the $\mathbf{L}$ from the previous slide.

Proof of Jaffe's Pumping Lemma

If the dfa has $\mathbf{k}$ states, then one can find among the first $\mathbf{k+1}$ states two states which are the same. So let $\mathbf{x}$ and $\mathbf{xy}$ be the parts of the word $\mathbf{u}$ read when reading these states. One can easily see that the state at the end after parsing $\mathbf{xy^hz}$ is the same as after parsing $\mathbf{u}$ and therefore the derivatives $\mathbf{L_{xy^hz}}$ and $\mathbf{L_u}$ coincide.

For the other way round, note that one finds for each $\mathbf{u}$ of length $\mathbf{k}$ or more a shorter $\mathbf{v}$ with $\mathbf{L_u = L_v}$ and thus all derivatives equal to a derivative of a word $\mathbf{w}$ with $|\mathbf{w}| < \mathbf{k}$. So there are only finitely many derivatives and the language is regular.

Block Pumping Lemma

There are constants $\mathbf{k}, \mathbf{h}$ such that, given a word $\mathbf{u}$ of length at least $\mathbf{h}$ (say in $\{\mathbf{0}, \mathbf{1}\}^*$) with exactly $\mathbf{k}$ special symbols (say $@$) inserted such that between any two $@$ there is at least one normal symbol from $\mathbf{u}$, one can delete all but two of the $@\mathbf{s}$ from the word obtaining $\mathbf{x}@\mathbf{y}@\mathbf{z}$ such that $\mathbf{L}(\mathbf{xyz}) = \mathbf{L}(\mathbf{xy}^\ell\mathbf{z})$ for all ℓ.

The weak block pumping lemma requires in addition that the first block border is before the beginning of the word and the last one after the end of the word.

There are one-sided and two-sided versions of this block pumping lemma and also its weak form.

Block Pumping and Matching

Theorem [Ehrenfeucht, Parikh and Rozenberg 1981]
A language is regular iff it satisfies the two-sided block pumping lemma iff it satisfies the two-sided block pumping lemma with the extra condition that one only pumps down (omits $\mathbf{y}$).

Theorem [Varricchio 1997]
A language is regular iff it satisfies the two-sided block pumping lemma with the additional constraint that one can only pump up.

Example [Ehrenfeucht and Rozenberg 1983, Ross and Winklmann 1982]
The language of square-containing ternary words (of form $\mathbf{xyyz}$ with $\mathbf{y} \neq \varepsilon$) is not context-free. It satisfies those one-sided pumping lemmas which only allow to pump up.

Thus a matching pumping lemma must be two-sided or allow to pump down.

One-Sided Block Pumping

Example [Chak, Freivalds, Stephan and Tan 2016]
The language of all binary words which are cube-containing or have a length which is not a power of 10 is one-sided block pumpable.
The language of all ternary words which are square-containing or have a length which is not a power of 10 is one-sided block pumpable.

Thus the one-sided block pumping lemma is not matching.

Theorem
Only a subclass of the regular languages satisfies the one-sided block pumping lemma with constant $k = 2$ and arbitrary h. These languages are unions of a finite set with sets of the form $\{w \in \Upsilon^* : |w| \geq h\}$ where Υ is a subset of the alphabet.

Computing the pump

An automatic function is a function recognised by a finite automaton which reads input and output synchronously, that is, each of them with one symbol per cycle with a special symbol for the case that input or output are exhausted.

So a pump inside a word, say 010011010, will be marked off by inserting $@$ at the beginning and end, the output is then $01@0011@010$ and the pump is 0011.

Thus one can simplify the concept of automatic function here with saying that an NFA goes over the word and is in special states while marking off the pump.

Regular Ogden's Lemma

Ogden's Lemma allows to mark some symbols by making them green. One could consider it as a game.

Ogden's Pumping Lemma for Regular Languages
A language $\mathbf{L}$ satisfies the lemma iff there is a constant $\mathbf{k}$ such that for every input word $\mathbf{w}$ out of which at least $\mathbf{k}$ symbols are green, the word can be split into $\mathbf{x}, \mathbf{y}, \mathbf{z}$ such that $\mathbf{y}$ contains at least one and at most $\mathbf{k}$ green symbols and $\mathbf{L(xyz) = L(xy^hz)}$ for all $\mathbf{h} \in \mathbb{N}$.

Example. $\mathbf{L}$ is all binary words with an even number of $\mathbf{1}$.
Constant is $\mathbf{k = 2}$.
If there is a green $\mathbf{0}$ in the input word then pump this $\mathbf{0}$ else there are two green $\mathbf{1}$ and pump an interval from one green $\mathbf{1}$ to the next $\mathbf{1}$ before or after it including the $\mathbf{0}$s in between.
So $\mathbf{001001001}$ is mapped to $\mathbf{001@0@01001}$
and $\mathbf{0011001001}$ is mapped to $\mathbf{001@1001@001}$.

Two-Sided Automatic Ogden Lemma 1

Theorem
If a language L satisfies the two-sided automatic Ogden Lemma then L is regular.

Proof. The nondeterministic algorithm just assumes that always the first k symbols are green and this type of input is enough to decide membership of L and it computes while processing the input word u from front to back with constant memory the output, such a model is equivalent to an NFA.

The memory consists of two words v, w of up to k symbols plus a symbol a plus a subset p of the state-set N of the NFA computing the pump plus another state s of N to track the current pump.

The idea is to simulate a concatenated version of successive down-pumping of the input word until less than k symbols are left.

Two-Sided Automatic Ogden Lemma 2

Nondetermistic Algorithm.

Whenever the state transition function δ has choices, it nondeterministically picks one of the choices.

1. Initialise $\mathbf{w} = \varepsilon$, $\mathbf{p} = \emptyset$, $\mathbf{s} = \mathbf{startstate}$.
2. If input is exhausted then goto 8 else Read $\mathbf{a}$.
3. Update $\mathbf{s} = \delta(\mathbf{s}, \mathbf{a})$.
4. Update $\mathbf{p}$ to $\{\delta(\mathbf{b}, \mathbf{a}) : \mathbf{b} \in \mathbf{p}\}$.
5. If $\mathbf{s}$ says "$\mathbf{a}$ not part of pump" then $\mathbf{w} = \mathbf{wa}$.
6. If $|\mathbf{w}| = \mathbf{k}$ then begin $\mathbf{p} = \mathbf{p} \cup \{\mathbf{s}\}$; let $\mathbf{s} = \delta(\mathbf{startstate}, \mathbf{w})$; let $\mathbf{v}$ be symbols of $\mathbf{w}$ after removal of those in the new pump (note that $|\mathbf{v}| < \mathbf{k}$); let $\mathbf{w} = \mathbf{v}$; end.
7. Goto 2.
8. If all states in $\mathbf{p} \cup \{\mathbf{s}\}$ are accepting then output $\mathbf{L}(\mathbf{w})$ else reject computation.

Applications

Above algorithm also works for two-sided traditional pumping lemma and two-sided weak block pumping lemma.

For the first, the pump is always among the first $\mathbf{k}$ symbols and covered by above algorithm.

For the latter, one makes all blocks except the last one containing one symbol and then the pump is either a subset of the first $\mathbf{k}$ symbols or a tail starting somewhere among the first $\mathbf{k+1}$ symbols until the end of the word. This is also covered by the above algorithm.

Theorem

Thus the automatic versions of the two-sided regular Ogden's Pumping Lemma, the two-sided traditional pumping lemma and the two-sided weak block pumping lemma are all matching.

Regular Ogden Lemma 1

Proposition
Let $\mathbf{f(x)}$ be the number of $\mathbf{1}$ minus the number of $\mathbf{0}$ in the word $\mathbf{x}$ and $\mathbf{L = \{x \in \{0, 1, \ldots\}^* : f(x) = 0\}}$. Now $\mathbf{L}$ satisfies the two-sided regular Ogden Lemma.

Proof. The constant is $\mathbf{2}$. Given a word $\mathbf{u}$ with at least two green characters, one splits the word into blocks such that each block contains exactly one green letter. Now one takes the first option which applies:

If there is a block $\mathbf{y}$ with $\mathbf{f(y) = 0}$ then one pumps this block. Pumping a balanced word does not change membership in $\mathbf{L}$ as $\mathbf{f(xy^kz) = f(xyz)}$ for all $\mathbf{k} \in \mathbb{N}$.

Regular Ogden Lemma 2

If there are two neighbouring blocks $\mathbf{y}, \mathbf{z}$ with $\mathbf{f}(\mathbf{y}) \cdot \mathbf{f}(\mathbf{z}) < \mathbf{0}$ then one selects first that block, say $\mathbf{y}$ with $\mathbf{f}(\mathbf{y}) \cdot \mathbf{f}(\mathbf{yz}) \leq \mathbf{0}$. Let $\mathbf{z}' = \mathbf{z}$ and as long as $\mathbf{f}(\mathbf{yz}') \neq \mathbf{0}$ one deletes the last symbol of $\mathbf{z}'$. This algorithm terminates with nonempty $\mathbf{z}'$ due to the discrete intermediate value theorem. Now one pumps $\mathbf{yz}'$.

The remaining case is that all blocks have the same sign, say $\mathbf{f}(\mathbf{y}) > \mathbf{0}$ for all blocks $\mathbf{y}$ in the splitting. So one picks any $\mathbf{y}$ and pumps it, all resulting words $\mathbf{xy^kz}$ satisfy $\mathbf{f}(\mathbf{xy^kz}) > \mathbf{0}$. Thus pumping is also here possible.

Note. $\mathbf{L}$ does not satisfy the one-sided regular traditional pumping lemma.

Two-Sided Pumping Limitations

Theorem
The automatic two-sided weak pumping lemma is satisfied by $\{u \in \{0,1\}^* : u$ has as many 0s as 1s$\}$.

Method: Pump first occurrence of $01, 10$ which exists; for $\{a\}^+\{a\}^+$, pump first symbol. Constant is $k = 2$.

However, every language L which obeys the automatic weak pumping lemma satisfies that it can be decided in quadratic time.

One-sided Automatic Pumping

Theorem
The one-sided automatic regular Ogden Lemma is not matching.

Method: Let $\mathbf{L}$ be the language of all square-containing ternary words and let $\mathbf{k = 20}$. If there is a square of up to $\mathbf{18}$ symbols then pump any marked symbol outside it else there are two occurrences of a symbol $\mathbf{a}$ with a marked $\mathbf{b}$ plus perhaps up to two further symbols between the two $\mathbf{a}$, pump the symbols strictly between the two $\mathbf{a}$.

The same applies to traditional pumping lemma and weak traditional pumping lemma.

Fact [Chak, Freivalds, Stephan and Tan 2016]
All languages of the form $\mathbf{L}^*$ satisfy the one-sided weak block pumping lemma and the pump is the full word, thus automatic.

Regular Bader-Moura Lemma

Definition [Bader and Moura 1982]
The regular version of the Pumping Lemma of Bader and Moura can be stated as follows: If $\mathbf{L}$ is regular then there is a constant $\mathbf{k}$ such that for all words $\mathbf{u}$ which have distinguished symbols (painted in green) and excluded symbols (painted in red) where the word contains a block (subword) having at exactly $\mathbf{k}$ green and no red symbols, one can split $\mathbf{u}$ into $\mathbf{x}, \mathbf{y}, \mathbf{z}$ such that $\forall \mathbf{h} \in \mathbb{N}\,[\mathbf{L}(\mathbf{u}) = \mathbf{L}(\mathbf{x}\mathbf{y}^{\mathbf{h}}\mathbf{z})]$ and in $\mathbf{y}$ are $\mathbf{1}$ to $\mathbf{k}$ green and no red symbols.

Note that this subword condition is satisfied whenever the number of the green symbols is at least $(\mathbf{k}-\mathbf{1})$ times the red symbols plus $\mathbf{k}$ and that for the context-free original version the green symbols had to outnumber the red ones exponentially.

The two-sided automatic Bader-Moura Lemma is matching, so the one-sided is of interest.

Example

Let the alphabet contain at least two symbols and let $\mathbf{L}$ contain all cube-containing words as well as all words whose length is not a nonzero power of $\mathbf{10}$. Then $\mathbf{L}$ satisfies the automatic Bader-Moura lemma with constant $\mathbf{2}$.

Given a word $\mathbf{u}$ with a subword containing two green and no red symbols, one can split $\mathbf{u}$ as $\mathbf{vabw}$ where $\mathbf{a}$ is a green symbol and $\mathbf{b}$ is a further symbol which is not red next to $\mathbf{a}$.

If the length of $\mathbf{u}$ is odd then all words $\mathbf{v(ab)^h w}$ with $\mathbf{h} \in \{\mathbf{0}\} \cup \{\mathbf{2},\mathbf{3},\mathbf{4},\ldots\}$ are in $\mathbf{L}$ else all words $\mathbf{va^h bw}$ with $\mathbf{h} \in \{\mathbf{0}\} \cup \{\mathbf{2},\mathbf{3},\mathbf{4},\ldots\}$ are in $\mathbf{L}$. This is due to the fact that $\mathbf{h} \in \{\mathbf{0},\mathbf{2}\}$ produces an odd length word and $\mathbf{h} \geq \mathbf{3}$ produces a cube-containing word.

Note that an automatic function can detect whether the length is odd and where the first occurrence of an $\mathbf{ab}$ as above in the word is.

Generalisation

Theorem
If $\mathbf{L}$ is (automatically) one-sided block pumpable then $\mathbf{L}$ satisfies (automatically) the one-sided regular pumping Lemma of Bader and Moura.

Proof. Let $\mathbf{k}$ be the block pumping constant and choose the same constant $\mathbf{k}$ for the Bader-Moura Lemma.

Given now a word $\mathbf{u} \in \mathbf{L}$ with red and green symbols such that there are $\mathbf{k}$ green symbols with no red ones between them, an automatic function can now insert the separation markers $@$ exactly before each of the $\mathbf{k}$ green symbols and split the word $\mathbf{u}$ into blocks so that each inner block contains exactly one green and no red symbol. Now the pump consists of some consecutive inner blocks.

If the pump selection function of block pumping lemma is automatic, also the overall function to compute the pump from the coloured word is automatic.

Open Problem

Conjecture

If a language L is automatically one-sided block pumpable then it is regular.

Note that for the case of only one inner block, the block pumping function is trivial and thus automatic and every language pumpable with this constant is regular.

Furthermore, for the unary alphabet, one-sided block pumpable languages are regular. This holds in general for all one-sided block pumpable languages with polynomial growth [Chak, Freivalds, Stephan, Tan 2016].

None of the known examples of one-sided block pumpable languages admits an automatic pump function.

Summary

Two-sided automatic pumping is matching, that is, characterising the regular languages, for the traditional pumping lemma, the regular variant of Ogden's Pumping Lemma, the weak block pumping lemma as well as all pumping lemmas stronger than these.

The automatic two-sided weak traditional pumping lemma is not matching, but is satisfied only by quadratic time decidable languages.

For one-sided pumping, various examples are found where one-sided automatic pumping is not matching, for example Ogden's Pumping Lemma and traditional pumping lemma and weak block pumping lemma.

Open Problem. Is the one-sided automatic block pumping lemma matching?