Algebraic theory of automata: historical perspective and new advances (Part II)

Jean-Éric Pin[1]

[1]LIAFA, CNRS and Université Paris Diderot

Warsaw, 2011

Summary

(1) Profinite topology on A^*
(2) Equational theories for lattices of languages
(3) Some examples
(4) Profinite topologies

A reminder on metric spaces

A metric space is a set E equipped with a metric d.

A sequence $(x_n)_{n\geqslant 0}$ is Cauchy if for each $\varepsilon > 0$, there exists k such that, for each $n \geqslant k$ and $m \geqslant k$, $d(x_n, x_m) < \varepsilon$.

A function φ from (E, d) into (E', d') is uniformly continuous if for each $\varepsilon > 0$, there exists $\delta > 0$ such that $d(x, y) < \delta$ implies $d'(\varphi(x), \varphi(y)) < \varepsilon$.

A metric space is complete if every Cauchy sequence is convergent.

Completion of a metric space

A completion of a metric space E is a complete metric space $\widehat{E}$ together with an isometric embedding of E as a dense subspace of $\widehat{E}$.

Every metric space admits a completion, which is unique up to uniform isomorphism. For instance, the completion of $\mathbb{Q}$ is $\mathbb{R}$.

Any uniformly continuous function $\varphi : E \to E'$ admits a unique uniformly continuous extension $\hat{\varphi} : \widehat{E} \to \widehat{E'}$.

Two examples

Let E be a finite set. The discrete metric d is defined by $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ otherwise. Then (E, d) is a complete metric space.

Let p be a prime number. The p-adic valuation of a non-zero integer n is

$$\nu_p(n) = \max\left\{k \in \mathbb{N} \mid p^k \text{ divides } n\right\}$$

By convention, $\nu_p(0) = +\infty$. The p-adic norm of n is the real number $|n|_p = p^{-\nu_p(n)}$. Finally, the metric d_p is defined by $d_p(u, v) = |u - v|_p$. The completion of $\mathbb{N}$ for d_p is the set of p-adic numbers.

Part I

The profinite world

Citation (M. Stone)

A cardinal principle of modern mathematical research may be stated as a maxim: One must always topologize.

Separating words

A deterministic finite automaton (DFA) separates two words if it accepts one of the words but not the other one.

A monoid M separates two words u and v of A^* if there exists a monoid morphism $\varphi : A^* \to M$ such that $\varphi(u) \neq \varphi(v)$.

Proposition

One can always separate two distinct words by a finite automaton (respectively by a finite monoid).

Separating words

- The morphism which maps each word onto its length modulo 2 is a morphism from $\{a, b\}^*$ onto $\mathbb{Z}/2\mathbb{Z}$ which separates $abaaba$ and $abaabab$.

- Similarly, for each letter a, one can count the number of a modulo n.

- Let $M = \{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right), \left(\begin{smallmatrix}1&0\\1&0\end{smallmatrix}\right), \left(\begin{smallmatrix}0&1\\0&1\end{smallmatrix}\right)\}$ and let $\varphi : \{a, b\}^* \to M$ defined by $\varphi(a) = \left(\begin{smallmatrix}1&0\\1&0\end{smallmatrix}\right)$ and $\varphi(b) = \left(\begin{smallmatrix}0&1\\0&1\end{smallmatrix}\right)$. Then, for all u, φ separates ua and ub since $\varphi(ua) = \varphi(a)$ and $\varphi(ub) = \varphi(b)$.

The profinite metric

Let u and v be two words. Put

$$r(u,v) = \min\{|M| \mid M \text{ is a finite monoid that separates } u \text{ and } v\}$$

$$d(u,v) = 2^{-r(u,v)}$$

Then d is an ultrametric, that is, for all $x, y, z \in A^*$,

(1) $d(x,x) = 0$,

(2) $d(x,y) = d(y,x)$,

(3) $d(x,z) \leqslant \max\{d(x,y), d(y,z)\}$

Another profinite metric

Let

$$r'(u,v) = \min\{\#\text{ states}(\mathcal{A}) \mid \mathcal{A} \text{ is a finite DFA separating } u \text{ and } v\}$$

$$d'(u,v) = 2^{-r'(u,v)}$$

The metric d' is uniformly equivalent to d:

$$2^{-\frac{1}{d'(u,v)}} \leqslant d(u,v) \leqslant d'(u,v)$$

Therefore, a function is uniformly continuous for d iff it is uniformly continuous for d'.

Main properties of d

Intuitively, two words are close for d if one needs a large monoid to separate them.

A sequence of words u_n is a Cauchy sequence iff, for every morphism φ from A^* to a finite monoid, the sequence $\varphi(u_n)$ is ultimately constant.

A sequence of words u_n converges to a word u iff, for every morphism φ from A^* to a finite monoid, the sequence $\varphi(u_n)$ is ultimately equal to $\varphi(u)$.

The free profinite monoid

The completion of the metric space (A^*, d) is the free profinite monoid on A and is denoted by $\widehat{A^*}$. It is a compact space, whose elements are called profinite words.

The concatenation product is uniformly continuous on A^* and can be extended by continuity to $\widehat{A^*}$.

Any morphism $\varphi : A^* \to M$, where M is a (discrete) finite monoid extends in a unique way to a uniformly continuous morphism $\hat{\varphi} : \widehat{A^*} \to M$.

The free profinite monoid as a projective limit

The monoid $\widehat{A^*}$ can be defined as the projective limit of the directed system formed by the surjective morphisms between finite A-generated monoids.

Let Φ be the class of all morphisms from A^* onto a finite monoid. Consider the product monoid

$$M = \prod_{\varphi \in \Phi} \varphi(A^*)$$

A family $(s_\varphi)_{\varphi \in \Phi}$ (where $s_\varphi \in \varphi(A^*)$) is compatible if, for each morphism $\pi : \varphi(A^*) \to \pi(\varphi(A^*))$, one has $s_{\pi \circ \varphi} = \pi(s_\varphi)$. Then $\widehat{A^*}$ is the submonoid of M formed by the compatible elements.

Profinite words

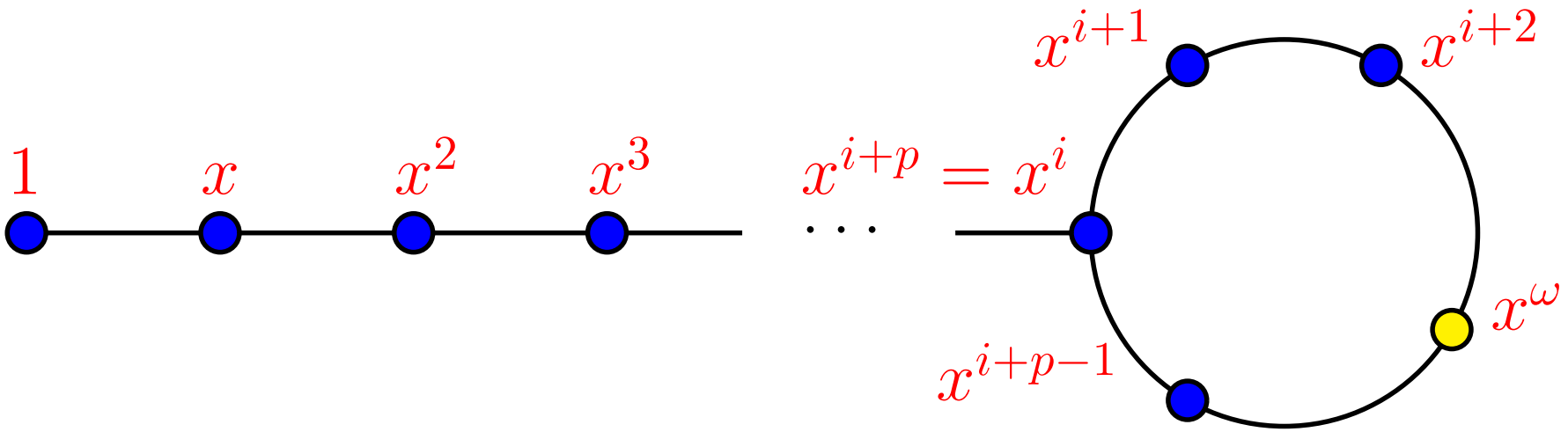
A profinite word u is completely determined by the elements $\hat{\varphi}(u)$, where φ runs over Φ.

$$\text{Profinite word } u \leftrightarrow \{\hat{\varphi}(u)\}_{\varphi \in \Phi}$$

Alternatively, one can define a profinite word as the limit of a Cauchy sequence of finite words, up to the following equivalence: two Cauchy sequences $x = (x_n)_{n \geqslant 0}$ and $y = (y_n)_{n \geqslant 0}$ are equivalent if the interleave sequence $x_0, y_0, x_1, y_1, \ldots$ is also a Cauchy sequence.

The profinite operator ω

For each $u \in A^*$, the sequence $u^{n!}$ is a Cauchy sequence and hence converges in $\widehat{A^*}$ to a limit, denoted by u^ω. If φ is a morphism from A^* onto a finite monoid, $\varphi(u^\omega)$ is the unique idempotent x^ω of the semigroup generated by $x = \varphi(u)$.



Another profinite word

Let us fix a total order on the alphabet A. Let $u_0, u_1, \ldots$ be the ordered sequence of all words of A^* in the induced shortlex order.

$$1, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, \ldots$$

Reilly and Zhang (see also Almeida-Volkov) proved that the sequence $(v_n)_{n\geqslant 0}$ defined by

$$v_0 = u_0, \ v_{n+1} = (v_n u_{n+1} v_n)^{(n+1)!}$$

is a Cauchy sequence, which converges to an idempotent ρ_A of the minimal ideal of $\widehat{A^*}$.

Regular languages and clopen sets

The maps $L \mapsto \overline{L}$ and $K \mapsto K \cap A^*$ are inverse isomorphisms between the Boolean algebras $\mathrm{Reg}(A^*)$ and $\mathrm{Clopen}(\widehat{A^*})$. For all regular langauges L, L_1, L_2 of A^*:

(1) $\overline{L^c} = (\overline{L})^c$,

(2) $\overline{L_1 \cup L_2} = \overline{L}_1 \cup \overline{L}_2$,

(3) $\overline{L_1 \cap L_2} = \overline{L}_1 \cap \overline{L}_2$,

(4) for all $x, y \in A^*$, then $\overline{x^{-1}Ly^{-1}} = x^{-1}\overline{L}y^{-1}$.

(5) If $\varphi : A^* \to B^*$ is a morphism and $L \in \mathrm{Reg}(B^*)$, then $\hat{\varphi}^{-1}(\overline{L}) = \overline{\varphi^{-1}(L)}$.

Part II

Equational theory for lattices

Lattices of languages

Let A be a finite alphabet. A lattice of languages is a set of regular languages of A^* containing $\emptyset$ and A^* and closed under finite intersection and finite union.

Let u and v be words of A^*. A language L of A^* satisfies the equation $u \longrightarrow v$ if

$$u \in L \Rightarrow v \in L$$

Let E be a set of equations of the form $u \longrightarrow v$. Then the languages of A^* satisfying the equations of E form a lattice of languages.

Equational description of finite lattices

Proposition

A finite set of languages of A^* is a lattice of languages iff it can be defined by a set of equations of the form $u \to v$ with $u, v \in A^*$.

Therefore, there is an equational theory for finite lattices of languages. What about infinite lattices?

One needs the profinite world...

Profinite equations

Let (u, v) be a pair of profinite words of $\widehat{A^*}$. We say that a regular language L of A^* satisfies the profinite equation $u \to v$ if

$$u \in \overline{L} \Rightarrow v \in \overline{L}$$

Let $\eta : A^* \to M$ be the syntactic morphism of L. Then L satisfies the profinite equation $u \to v$ iff

$$\hat{\eta}(u) \in \eta(L) \Rightarrow \hat{\eta}(v) \in \eta(L)$$

Equational theory of lattices

Given a set E of equations of the form $u \longrightarrow v$ (where u and v are profinite words), the set of all regular languages of A^* satisfying all the equations of E is called the set of languages defined by E.

Theorem (Gehrke, Grigorieff, Pin 2008)

A set of regular languages of A^* is a lattice of languages iff it can be defined by a set of equations of the form $u \longrightarrow v$, where $u, v \in \widehat{A^*}$.

Equations of the form $u \leqslant v$

Let us say that a regular language satisfies the equation $u \leqslant v$ if, for all $x, y \in \widehat{A^*}$, it satisfies the equation $xvy \to xuy$.

Proposition

Let L be a regular language of A^*, let $(M, \leqslant_L)$ be its syntactic ordered monoid and let $\eta : A^* \to M$ be its syntactic morphism. Then L satisfies the equation $u \leqslant v$ iff $\hat{\eta}(u) \leqslant_L \hat{\eta}(v)$.

Quotienting algebras of languages

A lattice of languages is a quotienting algebra of languages if it is closed under the quotienting operations $L \to u^{-1}L$ and $L \to Lu^{-1}$, for each word $u \in A^*$.

Theorem

A set of regular languages of A^* is a quotienting algebra of languages iff it can be defined by a set of equations of the form $u \leqslant v$, where $u, v \in \widehat{A^*}$.

Boolean algebras

Let us write

$u \leftrightarrow v$ for $u \longrightarrow v$ and $v \longrightarrow u$,

$u = v$ for $u \leqslant v$ and $v \leqslant u$.

Theorem

(1) *A set of regular languages of A^* is a Boolean algebra iff it can be defined by a set of equations of the form $u \leftrightarrow v$.*

(2) *It is a Boolean algebra closed under quotients iff it can be defined by a set of equations of the form $u = v$.*

Interpreting equations

Let u and v be two profinite words.

Closed under		**Interpretation**
$\cup, \cap$	$u \to v$	$u \in \overline{L} \Rightarrow v \in \overline{L}$
$+$ quotient	$u \leqslant v$	$\forall x, y \quad xvy \to xuy$
$+$ complement (L^c)	$u \leftrightarrow v$	$u \to v$ and $v \to u$
$+$ quotient and L^c	$u = v$	$xuy \leftrightarrow xvy$

Identities

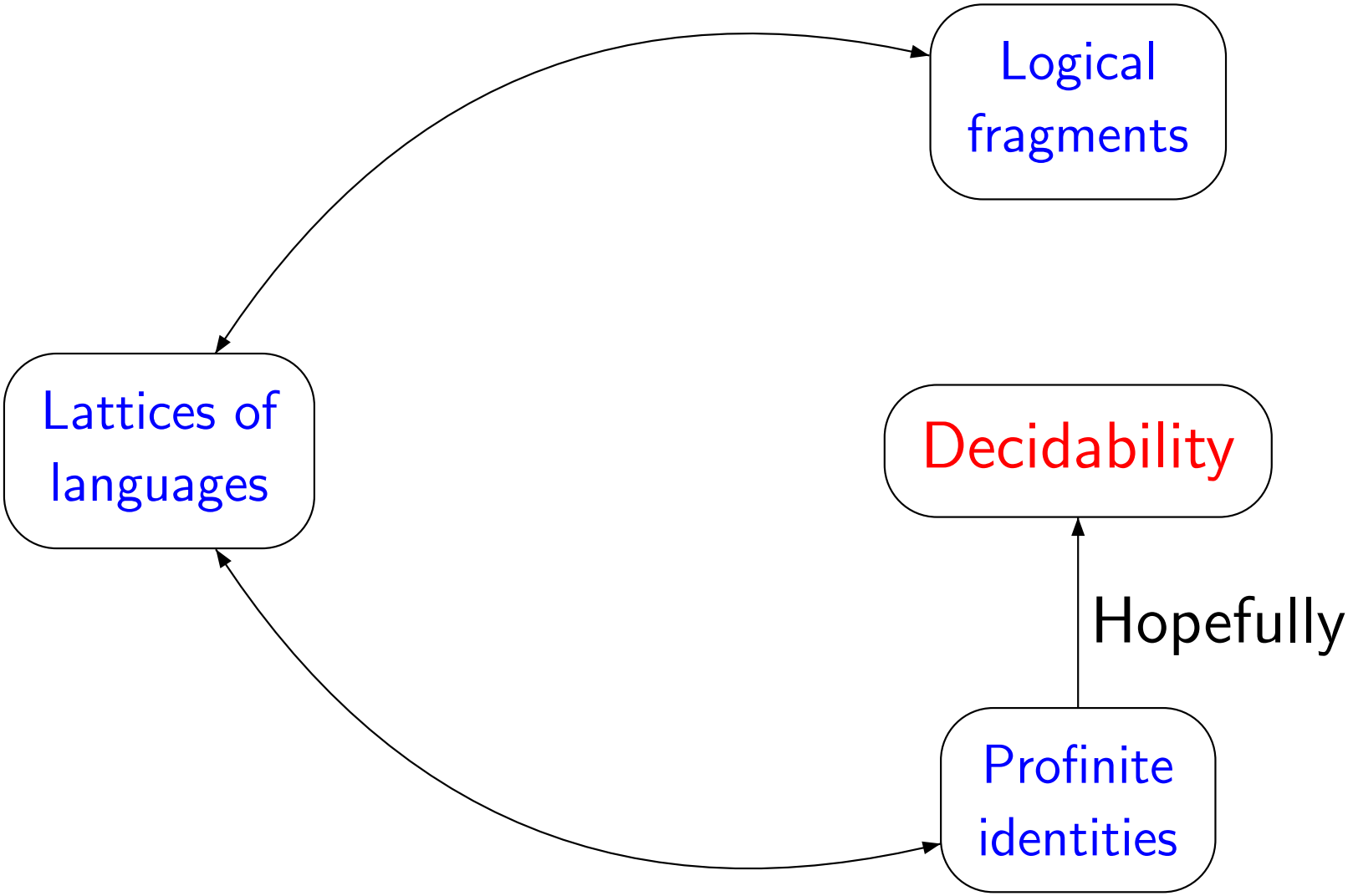
One can also recover Eilenberg's variety theorem and its variants by using identities. An identity is an equation in which letters are considered as variables.

Closed under inverse of $\cdots$ morphisms	**Interpretation of variables**
all	words
length increasing	nonempty words
length preserving	letters
length multiplying	words of equal length

Equational descriptions

- Every lattice of regular languages has an equational description.
- In particular, any class of regular languages defined by a fragment of logic closed under conjunctions and disjunctions (first order, monadic second order, temporal, etc.) admits an equational description.
- This result can also be adapted to languages of infinite words, words over ordinals or linear orders, and hopefully to tree languages.

The virtuous circle



Part III

Some examples

- Languages with zero
- Nondense languages
- Slender languages
- Sparse languages
- Examples from logic
- Examples of identities

Languages with zero

A language with zero is a language whose syntactic monoid has a zero. The class of regular languages with zero is closed under Boolean operations and quotients, but not under inverse of morphisms.

Proposition

A regular language has a zero iff it satisfies the equation $x\rho_A = \rho_A = \rho_A x$ *for all* $x \in A^*$.

In the sequel, we simply write 0 for ρ_A to mean that L has a zero.

Nondense languages

A language L of A^* is dense if, for each word $u \in A^*$, $L \cap A^* u A^* \neq \emptyset$.

Regular non-dense or full languages form a lattice closed under quotients.

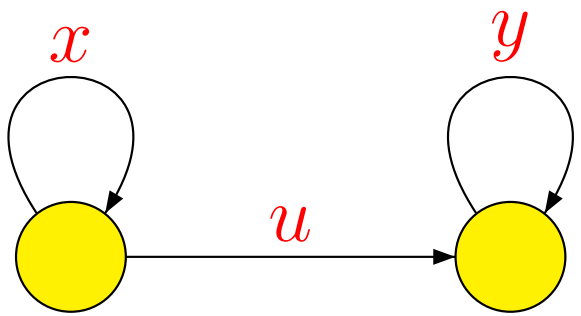
Theorem

A regular language of A^* is non-dense or full iff it satisfies the equations $x \leqslant 0$ for all $x \in A^*$.

Slender or full languages

A regular language is slender iff it is a finite union of languages of the form xu^*y, where $x, u, y \in A^*$.

Fact. A regular language is slender iff its minimal deterministic automaton does not contain any pair of connected cycles.



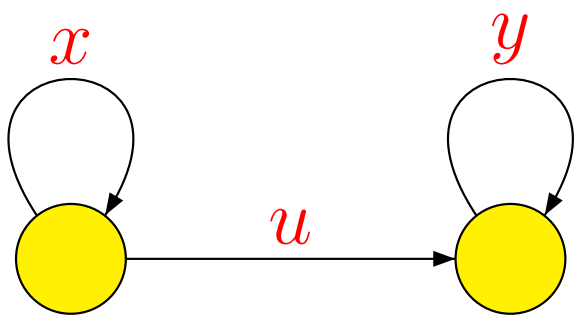
Two connected cycles, where $x, y \in A^+$ and $u \in A^*$.

Equations for slender languages

Denote by $i(x)$ the initial of a word x.

Theorem

Suppose that $|A| \geqslant 2$. A regular language of A^* is slender or full iff it satisfies the equations $x \leqslant 0$ for all $x \in A^*$ and the equation $x^\omega u y^\omega = 0$ for each $x, y \in A^+$, $u \in A^*$ such that $i(uy) \neq i(x)$.



Sparse languages

A regular language is sparse iff it is a finite union of languages of the form $u_0 v_1^* u_1 \cdots v_n^* u_n$, where u_0, v_1, $\ldots$, v_n, u_n are words.

Theorem

Suppose that $|A| \geqslant 2$. A regular language of A^* is sparse or full iff it satisfies the equations $x \leqslant 0$ for all $x \in A^*$ and the equations $(x^\omega y^\omega)^\omega = 0$ for each $x, y \in A^+$ such that $i(x) \neq i(y)$.

Boolean closures

Suppose that $|A| \geqslant 2$.

Theorem

A regular language of A^* is slender or coslender iff it satisfies the equations $x^\omega u y^\omega = 0$ for each $x, y \in A^+$, $u \in A^*$ such that $i(uy) \neq i(x)$.

Theorem

A regular language of A^* is sparse or cosparse iff it satisfies the equations $(x^\omega y^\omega)^\omega = 0$ for each $x, y \in A^+$ such that $i(x) \neq i(y)$.

Identities of well-known logical fragments

(1) Star-free languages: $x^{\omega+1} = x^\omega$. Captured by the logical fragment $FO[<]$.

(2) Finite unions of languages of the form $A^*a_1A^*a_2A^* \cdots a_kA^*$, where $a_1, \ldots, a_k$ are letters: $x \leqslant 1$. Captured by $\Sigma_1[<]$.

(3) Piecewise testable languages = Boolean closure of (3): $x^{\omega+1} = x^\omega$ and $(xy)^\omega = (yx)^\omega$. Captured by $\mathcal{B}\Sigma_1[<]$.

(4) Unambiguous star-free languages: $x^{\omega+1} = x^\omega$ and $(xy)^\omega(yx)^\omega(xy)^\omega = (xy)^\omega$. Captured by $FO_2[<]$ (first order with two variables) or by $\Sigma_2[<] \cap \Pi_2[<]$ or by unary temporal logic.

Another fragment of Büchi's sequential calculus

Denote by $\mathcal{B}\Sigma_1(S)$ the Boolean combinations of existential formulas in the signature $\{S, (\mathbf{a})_{a\in A}\}$. This logical fragment allows to specify properties like the factor aa occurs at least twice. Here is an equational description of the $\mathcal{B}\Sigma_1(S)$-definable languages, where $r, s, u, v, x, y \in A^*$:

$$ux^\omega v \leftrightarrow ux^{\omega+1}v$$
$$ux^\omega ry^\omega sx^\omega ty^\omega v \leftrightarrow ux^\omega ty^\omega sx^\omega ry^\omega v$$
$$x^\omega uy^\omega vx^\omega \leftrightarrow y^\omega vx^\omega uy^\omega$$
$$y(xy)^\omega \leftrightarrow (xy)^\omega \leftrightarrow (xy)^\omega x$$

Examples of length-multiplying identities

Length-multiplying identities: x and y represent words of the same length.

(1) Regular languages of AC^0: $(x^{\omega-1}y)^\omega = (x^{\omega-1}y)^{\omega+1}$. Captured by $FO[< +MOD]$.

(2) Finite union of languages of the form $(A^d)^*a_1(A^d)^*a_2(A^d)^* \cdots\ a_k(A^d)^*$, with $d > 0$: $x^{\omega-1}y \leqslant 1$ and $yx^{\omega-1} \leqslant 1$. Captured by $\Sigma_1[< +MOD]$.

Part IV

Profinite topologies

Profinite metrics (Boolean case)

Let $\mathcal{L}$ be a Boolean algebra of regular languages of A^*. A language L separates two words if it contains one of the words but not the other one. Put

$$r_{\mathcal{L}}(u,v) = \min\{\#(L) \mid L \text{ is a language of } \mathcal{L} \text{ that separates } u \text{ and } v\}$$

$$d_{\mathcal{L}}(u,v) = 2^{-r_{\mathcal{L}}(u,v)}$$

Intuitively, two words are close for $d_{\mathcal{L}}$ if one needs a complex language to separate them.

Properties of $d_{\mathcal{L}}$

For all $x, y, z \in A^*$,

(1) $d_{\mathcal{L}}(x, x) = 0$,

(2) $d_{\mathcal{L}}(x, y) = d_{\mathcal{L}}(y, x)$,

(3) $d_{\mathcal{L}}(x, z) \leqslant \max\{d_{\mathcal{L}}(x, y), d_{\mathcal{L}}(y, z)\}$

Thus $d_{\mathcal{L}}$ is a pseudo-ultrametric, which defines the pro-$\mathcal{L}$ topology. It is an ultrametric iff $\mathcal{L}$ separates words.

The completion of A^* for $d_{\mathcal{L}}$ is denoted by $\widehat{A^*}^{\mathcal{L}}$. It is a compact space (Hausdorff iff $\mathcal{L}$ separates words) and a monoid if $\mathcal{L}$ is closed under quotients.

Examples

If $\mathcal{L}$ finite or cofinite languages of A^*, then $\widehat{A^*}^{\mathcal{L}} = A^* \cup \{0\}$ (one point compactification).

If $\mathcal{L}$ is the set of languages of the form $FA^* \cup G$, where F and G are finite, then $\widehat{A^*}^{\mathcal{L}} = A^* \cup A^\omega$.

If $\mathcal{L}$ is the set of piecewise testable languages, then $\widehat{A^*}^{\mathcal{L}}$ is countable and is structure is well understood.

In general, it is a difficult problem to describe $\widehat{A^*}^{\mathcal{L}}$. See [J. Almeida, Gesammelte Werke].

$\mathcal{L}$-preserving functions

Definition. A function $f : A^* \to A^*$ is $\mathcal{L}$-preserving if, for each language $L \in \mathcal{L}$, $f^{-1}(L) \in \mathcal{L}$.

Theorem

A function from $f : A^* \to A^*$ is uniformly continuous for $d_{\mathcal{L}}$ iff it is $\mathcal{L}$-preserving.

- One can extend this result to lattices of regular languages by using quasi-uniform structures.

- Regular-preserving functions are exactly the uniformly continuous functions for d.

A well-known exercise...

If L is a language, its square root is $K = \{u \in A^* \mid u^2 \in L\}$.

Exercise. Show that the square root of a regular [star-free] language is regular [star-free].

Proof. Note that $K = f^{-1}(L)$, where $f(u) = u^2$. Let $\mathcal{L}$ be a quotienting algebra of languages. Since the product is uniformly continuous for $d_{\mathcal{L}}$, f is uniformly continuous. Thus f is $\mathcal{L}$-preserving.

Extension to lattices

If $\mathcal{L}$ is a lattice of languages, the same ideas can be applied. One defines a quasi-uniform structure generated by the sets

$$U_L = (L \times A^*) \cup (A^* \times L^c) \qquad (L \in \mathcal{L})$$

called the pro-$\mathcal{L}$ (quasi)-uniform structure.

Theorem

A function from A^* to A^* is uniformly continuous for the pro-$\mathcal{L}$ uniform structure iff it is $\mathcal{L}$-preserving.

Summary

- Every lattice of regular languages admits an equational description, a result that subsumes Eilenberg's variety theorem and its extensions.
- In particular, any class of regular languages defined by a fragment of logic closed under conjunctions and disjunctions (first order, monadic second order, temporal logic, etc.) admits an equational description.

Conclusion

Two difficult problems:

(1) Finding a set of equations defining a lattice can be difficult. In good cases, equations involve only words and simple operators like ω, but this is not the rule.

(2) Given a set of equations, one still needs to decide whether a given regular language satisfies these equations.