

Polynomial Operators on Classes of Regular Languages

Ondřej Klíma and Libor Polák

Department of Mathematics and Statistics
Masaryk University Brno

CAI 2009

The **polynomial operator** assigns to each class of languages \mathcal{V} the class of all (positive) boolean combinations the languages of the form

$$L_0 a_1 L_1 a_2 \dots a_\ell L_\ell, \quad (*)$$

where A is an alphabet, $a_1, \dots, a_\ell \in A$, $L_0, \dots, L_\ell \in \mathcal{V}(A)$ (i.e. they are over A).

The resulting classes are denoted by $\text{PPol}\mathcal{V}$ and $\text{BPol}\mathcal{V}$, respectively.

In the **restricted** case we fix a natural number k and we allow only $\ell \leq k$ in $(*)$. We get the classes $\text{PPol}_k\mathcal{V}$ and $\text{BPol}_k\mathcal{V}$, respectively.

The **polynomial operator** assigns to each class of languages \mathcal{V} the class of all (positive) boolean combinations the languages of the form

$$L_0 a_1 L_1 a_2 \dots a_\ell L_\ell, \quad (*)$$

where A is an alphabet, $a_1, \dots, a_\ell \in A$, $L_0, \dots, L_\ell \in \mathcal{V}(A)$ (i.e. they are over A).

The resulting classes are denoted by **PPol** \mathcal{V} and **BPol** \mathcal{V} , respectively.

In the **restricted** case we fix a natural number k and we allow only $\ell \leq k$ in $(*)$. We get the classes **PPol** $_k\mathcal{V}$ and **BPol** $_k\mathcal{V}$, respectively.

The **polynomial operator** assigns to each class of languages \mathcal{V} the class of all (positive) boolean combinations the languages of the form

$$L_0 a_1 L_1 a_2 \dots a_\ell L_\ell, \quad (*)$$

where A is an alphabet, $a_1, \dots, a_\ell \in A$, $L_0, \dots, L_\ell \in \mathcal{V}(A)$ (i.e. they are over A).

The resulting classes are denoted by **PPol** \mathcal{V} and **BPol** \mathcal{V} , respectively.

In the **restricted** case we fix a natural number k and we allow only $\ell \leq k$ in $(*)$. We get the classes **PPol** $_k\mathcal{V}$ and **BPol** $_k\mathcal{V}$, respectively.

1. Let $\mathcal{T}(A) = \{\emptyset, A^*\}$ for each finite set A . Then $\text{PPol}\mathcal{T}$ is level 1/2 of the **Straubing-Thérien hierarchy** and $\text{BPol}\mathcal{T} = \mathcal{V}_1$ is level 1, i.e. the piecewise testable languages.

Result (Simon - 1972): Decidability of the membership problem for the class \mathcal{V}_1 .

Open problem: Decidability of the membership problem for the class $\text{BPol}\mathcal{V}_1 = \mathcal{V}_2$.

2. Let $\mathcal{S}^+(A)$ be the set of all finite unions of the languages of the form B^* , where $B \subseteq A$, for each finite set A .

Result (Pin, Straubing): $\text{BPol}\mathcal{S}^+ = \mathcal{V}_2$.

Open problem – reformulation:

Is it decidable whether a given regular language $L \subseteq A^*$ can be expressed as a boolean combination languages of the form $B_0^* a_1 B_1^* a_2 \dots a_\ell B_\ell^*$, where $a_1, \dots, a_\ell \in A, B_0, \dots, B_\ell \subseteq A$.

1. Let $\mathcal{T}(A) = \{\emptyset, A^*\}$ for each finite set A . Then $\text{PPol}\mathcal{T}$ is level 1/2 of the **Straubing-Thérien hierarchy** and $\text{BPol}\mathcal{T} = \mathcal{V}_1$ is level 1, i.e. the piecewise testable languages.

Result (Simon - 1972): Decidability of the membership problem for the class \mathcal{V}_1 .

Open problem: Decidability of the membership problem for the class $\text{BPol}\mathcal{V}_1 = \mathcal{V}_2$.

2. Let $\mathcal{S}^+(A)$ be the set of all finite unions of the languages of the form B^* , where $B \subseteq A$, for each finite set A .

Result (Pin, Straubing): $\text{BPol}\mathcal{S}^+ = \mathcal{V}_2$.

Open problem – reformulation:

Is it decidable whether a given regular language $L \subseteq A^*$ can be expressed as a boolean combination languages of the form $B_0^* a_1 B_1^* a_2 \dots a_\ell B_\ell^*$, where $a_1, \dots, a_\ell \in A, B_0, \dots, B_\ell \subseteq A$.

1. Let $\mathcal{T}(A) = \{\emptyset, A^*\}$ for each finite set A . Then $\text{PPol}\mathcal{T}$ is level 1/2 of the **Straubing-Thérien hierarchy** and $\text{BPol}\mathcal{T} = \mathcal{V}_1$ is level 1, i.e. the piecewise testable languages.

Result (Simon - 1972): Decidability of the membership problem for the class \mathcal{V}_1 .

Open problem: Decidability of the membership problem for the class $\text{BPol}\mathcal{V}_1 = \mathcal{V}_2$.

2. Let $\mathcal{S}^+(A)$ be the set of all finite unions of the languages of the form B^* , where $B \subseteq A$, for each finite set A .

Result (Pin, Straubing): $\text{BPol}\mathcal{S}^+ = \mathcal{V}_2$.

Open problem – reformulation:

Is it decidable whether a given regular language $L \subseteq A^*$ can be expressed as a boolean combination languages of the form $B_0^* a_1 B_1^* a_2 \dots a_\ell B_\ell^*$, where $a_1, \dots, a_\ell \in A$, $B_0, \dots, B_\ell \subseteq A$.

3. Let $\mathcal{S}(A)$ be the set of all finite unions of the languages of the form \overline{B} , where $B \subseteq A$, for each finite set A . Here \overline{B} is the set of all words over A containing exactly the letters from B .

4. Let m be a fixed natural number. Let $\mathcal{A}_m(A)$ be the set of all boolean combinations of the languages of the form $L(a, r) = \{u \in A^* \mid |u|_a \equiv r \pmod{m}\}$, where $a \in A$ and $0 \leq r < m$, for each finite set A .

Notice that the classes \mathcal{T} , \mathcal{S} , \mathcal{A}_m are boolean varieties and \mathcal{S}^+ is a positive variety.

3. Let $\mathcal{S}(A)$ be the set of all finite unions of the languages of the form \overline{B} , where $B \subseteq A$, for each finite set A . Here \overline{B} is the set of all words over A containing exactly the letters from B .

4. Let m be a fixed natural number. Let $\mathcal{A}_m(A)$ be the set of all boolean combinations of the languages of the form $L(a, r) = \{u \in A^* \mid |u|_a \equiv r \pmod{m}\}$, where $a \in A$ and $0 \leq r < m$, for each finite set A .

Notice that the classes \mathcal{T} , \mathcal{S} , \mathcal{A}_m are boolean varieties and \mathcal{S}^+ is a positive variety.

3. Let $\mathcal{S}(A)$ be the set of all finite unions of the languages of the form \overline{B} , where $B \subseteq A$, for each finite set A . Here \overline{B} is the set of all words over A containing exactly the letters from B .

4. Let m be a fixed natural number. Let $\mathcal{A}_m(A)$ be the set of all boolean combinations of the languages of the form $L(a, r) = \{u \in A^* \mid |u|_a \equiv r \pmod{m}\}$, where $a \in A$ and $0 \leq r < m$, for each finite set A .

Notice that the classes \mathcal{T} , \mathcal{S} , \mathcal{A}_m are boolean varieties and \mathcal{S}^+ is a positive variety.

A **boolean variety of languages** \mathcal{V} associates to every finite alphabet A a class $\mathcal{V}(A)$ of regular languages over A in such a way that

- $\mathcal{V}(A)$ is closed under finite unions, finite intersections and complements (in particular, $\emptyset, A^* \in \mathcal{V}(A)$),
- $\mathcal{V}(A)$ is closed under derivatives, i.e.
 $L \in \mathcal{V}(A)$, $u, v \in A^*$ implies
 $u^{-1}Lv^{-1} = \{w \in A^* \mid uwv \in L\} \in \mathcal{V}(A)$,
- \mathcal{V} is closed under inverse morphisms, i.e.
 $f: B^* \rightarrow A^*$, $L \in \mathcal{V}(A)$ implies
 $f^{-1}(L) = \{v \in B^* \mid f(v) \in L\} \in \mathcal{V}(B)$.

To get the notion of a **positive** variety of languages, we use in the first item only intersections and unions (not complements).

A **boolean variety of languages** \mathcal{V} associates to every finite alphabet A a class $\mathcal{V}(A)$ of regular languages over A in such a way that

- $\mathcal{V}(A)$ is closed under finite unions, finite intersections and complements (in particular, $\emptyset, A^* \in \mathcal{V}(A)$),
- $\mathcal{V}(A)$ is closed under derivatives, i.e.
 $L \in \mathcal{V}(A)$, $u, v \in A^*$ implies
 $u^{-1}Lv^{-1} = \{w \in A^* \mid uwv \in L\} \in \mathcal{V}(A)$,
- \mathcal{V} is closed under inverse morphisms, i.e.
 $f: B^* \rightarrow A^*$, $L \in \mathcal{V}(A)$ implies
 $f^{-1}(L) = \{v \in B^* \mid f(v) \in L\} \in \mathcal{V}(B)$.

To get the notion of a **positive** variety of languages, we use in the first item only intersections and unions (not complements).

A **pseudovariety** of finite (ordered) monoids is a class of finite monoids closed under submonoids, morphic images and products of finite families. Similarly for ordered monoids. When defining a **variety** of (ordered) monoids we use arbitrary products.

The pseudovarieties of ordered monoids can be characterized by pseudoidentities. The pseudovarieties we consider here are **equational** – they are given by identities, or equivalently, they are of the form $\text{Fin } \mathbf{V}$ where \mathbf{V} is a variety of (ordered) monoids.

A **pseudovariety** of finite (ordered) monoids is a class of finite monoids closed under submonoids, morphic images and products of finite families. Similarly for ordered monoids. When defining a **variety** of (ordered) monoids we use arbitrary products.

The pseudovarieties of ordered monoids can be characterized by pseudoidentities. The pseudovarieties we consider here are **equational** – they are given by identities, or equivalently, they are of the form $\text{Fin } \mathbf{V}$ where \mathbf{V} is a variety of (ordered) monoids.

For a regular language $L \subseteq A^*$, we define the relations \sim_L and \preceq_L on A^* as follows: for $u, v \in A^*$ we have

$u \sim_L v$ if and only if $(\forall p, q \in A^*) (puq \in L \iff pvq \in L)$,

$u \preceq_L v$ if and only if $(\forall p, q \in A^*) (pvq \in L \implies puq \in L)$.

The relation \sim_L is the **syntactic congruence** of L on A^* . It is of **finite index** (i.e. there are finitely many classes), the quotient structure $M(L) = A^*/\sim_L$ is called the **syntactic monoid** of L .

The relation \preceq_L is the **syntactic quasiorder** of L and we have $\preceq_L \cap \succeq_L = \sim_L$. Hence \preceq_L induces an order on $M(L) = A^*/\sim_L$, namely: $u \sim_L \leq v \sim_L$ if and only if $u \preceq_L v$. We speak about the **syntactic ordered monoid** of L ; we denote the structure by $O(L)$.

For a regular language $L \subseteq A^*$, we define the relations \sim_L and \preceq_L on A^* as follows: for $u, v \in A^*$ we have

$u \sim_L v$ if and only if $(\forall p, q \in A^*) (puq \in L \iff pvq \in L)$,

$u \preceq_L v$ if and only if $(\forall p, q \in A^*) (pvq \in L \implies puq \in L)$.

The relation \sim_L is the **syntactic congruence** of L on A^* . It is of **finite index** (i.e. there are finitely many classes), the quotient structure $M(L) = A^*/\sim_L$ is called the **syntactic monoid** of L .

The relation \preceq_L is the **syntactic quasiorder** of L and we have $\preceq_L \cap \succeq_L = \sim_L$. Hence \preceq_L induces an order on $M(L) = A^*/\sim_L$, namely: $u \sim_L \leq v \sim_L$ if and only if $u \preceq_L v$. We speak about the **syntactic ordered monoid** of L ; we denote the structure by $O(L)$.

Result (Eilenberg, Pin)

Boolean varieties (positive varieties) of languages correspond to pseudovarieties of finite monoids (ordered monoids). The correspondence, written $\mathcal{V} \longleftrightarrow \mathbf{V}$ ($\mathcal{P} \longleftrightarrow \mathbf{P}$), is given by the following relationship: for $L \subseteq A^$ we have*

$L \in \mathcal{V}(A)$ if and only if $M(L) \in \mathbf{V}$

($L \in \mathcal{P}(A)$ if and only if $O(L) \in \mathbf{P}$).

Pseudovarieties of (ordered) monoids corresponding to the classes $\mathcal{T}, \mathcal{S}^+, \mathcal{S}, \mathcal{A}_m$ consist exactly of all finite members of the following varieties:

$$\mathbf{T} = \text{Mod}(x = y), \quad \mathbf{S}^+ = \text{Mod}(x^2 = x, xy = yx, 1 \leq x),$$

$$\mathbf{S} = \text{Mod}(x^2 = x, xy = yx), \quad \mathbf{A}_m = \text{Mod}(xy = yx, x^m = 1).$$

The names for the (ordered) monoids of the pseudovarieties $\mathbf{T}, \mathbf{S}^+, \mathbf{S}, \mathbf{A}_m$ are **trivial monoids (semilattices with the smallest element 1, semilattices and abelian groups of index m , respectively)**

Let $X = \{x_1, x_2, \dots\}$. A relation γ on X^* is **a finite characteristic** if it satisfies the following conditions:

- (i) γ is a quasiorder on X^* ;
- (ii) γ is compatible with the multiplication, i.e. for each $u, v, w \in X^*$ we have

$$u \gamma v \quad \text{implies} \quad uw \gamma vw, \quad wu \gamma wv ;$$

- (iii) γ is fully invariant, i.e. for each morphism $\varphi : X^* \rightarrow X^*$ and each $u, v \in X^*$ we have

$$u \gamma v \quad \text{implies} \quad \varphi(u) \gamma \varphi(v) ;$$

- (iv) for each finite subset Y of the set X , the set Y^* intersects only finitely many classes of $X^* / \gamma \cap \gamma^{-1}$.

A relation γ on X^* satisfying conditions (i) – (iii) is called a **fully invariant compatible quasiorder**. It determines a variety $\mathbf{V}_\gamma = \text{Mod } \gamma$ of ordered monoids. The operators Id and Mod are mutually inverse bijections between varieties of ordered monoids and fully invariant compatible quasiorders on X^* . Condition (iv) says that the finitely generated free ordered monoids in \mathbf{V}_γ are finite, i.e. \mathbf{V}_γ is **locally finite**.

The pseudovariety $\text{Fin } \mathbf{V}_\gamma$ of all finite members from \mathbf{V}_γ corresponds to the positive variety \mathcal{V}_γ of languages by

$$L \in \mathcal{V}_\gamma(A) \text{ if and only if } \text{O}(L) \in \text{Fin } \mathbf{V}_\gamma, \text{ for all finite } A.$$

A relation γ on X^* satisfying conditions (i) – (iii) is called a **fully invariant compatible quasiorder**. It determines a variety $\mathbf{V}_\gamma = \text{Mod } \gamma$ of ordered monoids. The operators Id and Mod are mutually inverse bijections between varieties of ordered monoids and fully invariant compatible quasiorders on X^* . Condition (iv) says that the finitely generated free ordered monoids in \mathbf{V}_γ are finite, i.e. \mathbf{V}_γ is **locally finite**.

The pseudovariety $\text{Fin } \mathbf{V}_\gamma$ of all finite members from \mathbf{V}_γ corresponds to the positive variety \mathcal{V}_γ of languages by

$$L \in \mathcal{V}_\gamma(A) \text{ if and only if } \text{O}(L) \in \text{Fin } \mathbf{V}_\gamma, \text{ for all finite } A .$$

Finite characteristic of classes of languages

We say that γ is a **finite characteristic of a class of languages** \mathcal{V} if γ is a finite characteristic and for each finite alphabet A :

$$L \in \mathcal{V}(A) \quad \text{if and only if} \quad \gamma_A \subseteq \sim_L .$$

(Here A can be identified with a subset of X and γ_A is the restriction of γ to A^* .)

Lemma. Let \mathcal{V} be a class of languages and γ be a finite characteristic of \mathcal{V} . Then:

- (i) \mathcal{V} is equal to the positive variety of languages \mathcal{V}_γ ;
- (ii) if \mathbf{V} is the pseudovariety of finite ordered monoids corresponding to \mathcal{V} then $\gamma = \text{Id}\langle \mathbf{V} \rangle$.

The classes of languages in our basic examples have the following finite characteristics:

1. $\text{Id}\mathbf{T} = X^* \times X^*$.
2. $\text{Id}\mathbf{S}^+ = \{(u, v) \in X^* \times X^* \mid c(u) \subseteq c(v)\}$.
3. $\text{Id}\mathbf{S} = \{(u, v) \in X^* \times X^* \mid c(u) = c(v)\}$.
4. $\text{Id}\mathbf{A}_m = \{(u, v) \in X^* \times X^* \mid (\forall x \in X) |u|_x \equiv |v|_x \pmod{m}\}$.

Finite characteristic of classes of languages

We say that γ is a **finite characteristic of a class of languages** \mathcal{V} if γ is a finite characteristic and for each finite alphabet A :

$$L \in \mathcal{V}(A) \quad \text{if and only if} \quad \gamma_A \subseteq \sim_L .$$

(Here A can be identified with a subset of X and γ_A is the restriction of γ to A^* .)

Lemma. Let \mathcal{V} be a class of languages and γ be a finite characteristic of \mathcal{V} . Then:

- (i) \mathcal{V} is equal to the positive variety of languages \mathcal{V}_γ ;
- (ii) if \mathbf{V} is the pseudovariety of finite ordered monoids corresponding to \mathcal{V} then $\gamma = \text{Id} \langle \mathbf{V} \rangle$.

The classes of languages in our basic examples have the following finite characteristics:

1. $\text{Id} \mathbf{T} = X^* \times X^*$.
2. $\text{Id} \mathbf{S}^+ = \{ (u, v) \in X^* \times X^* \mid c(u) \subseteq c(v) \}$.
3. $\text{Id} \mathbf{S} = \{ (u, v) \in X^* \times X^* \mid c(u) = c(v) \}$.
4. $\text{Id} \mathbf{A}_m = \{ (u, v) \in X^* \times X^* \mid (\forall x \in X) |u|_x \equiv |v|_x \pmod{m} \}$.

Proposition

Let \mathcal{V} be a positive variety of languages and let \mathbf{V} be the corresponding pseudovariety of ordered monoids. Then the following conditions are equivalent.

- (i) For each finite alphabet A , the set $\mathcal{V}(A)$ is finite.*
- (ii) The pseudovariety of ordered monoids \mathbf{V} is locally finite, i.e. each finitely generated submonoid of an arbitrary product of ordered monoids from \mathbf{V} is finite.*
- (iii) There exists a finite characteristic of \mathcal{V} .*

Let k be a fixed natural number and γ be a finite characteristic.
For a word $u \in X^*$, we say that

$$f = (u_0, a_1, \dots, a_\ell, u_\ell)$$

is a **factorization** of u of length ℓ if $u_0, u_1, \dots, u_\ell \in X^*$,
 $a_1, a_2, \dots, a_\ell \in X$ and $u_0 a_1 u_1 \dots a_\ell u_\ell = u$.

The set of all factorizations of lengths at most k of the word u is
denoted by $\text{Fact}_k(u)$.

Main construction

For a factorization $f = (u_0, a_1, u_1, \dots, a_\ell, u_\ell)$ of a word $u \in X^*$ and a factorization $g = (v_0, b_1, v_1, \dots, b_m, v_m)$ of a word $v \in X^*$, we write $f \leq_\gamma g$ if

- $\ell = m$,
- $a_i = b_i$ for every $i \in \{1, \dots, \ell\}$,
- $u_i \gamma v_i$ for every $i \in \{0, 1, \dots, \ell\}$.

We define the relation $\mathbf{p}_k(\gamma)$ on the set X^* as follows:
for $u, v \in X^*$, we have $(u, v) \in \mathbf{p}_k(\gamma)$ iff

$$(\forall g \in \text{Fact}_k(v)) (\exists f \in \text{Fact}_k(u)) f \leq_\gamma g .$$

Theorem

Let \mathcal{V} be a locally finite positive variety of languages and γ be a finite characteristic of \mathcal{V} . Then $\text{PPol}_k \mathcal{V}$ is a locally finite positive variety of languages with the finite characteristic $\mathfrak{p}_k(\gamma)$ and $\text{BPol}_k \mathcal{V}$ is a locally finite boolean variety of languages with the finite characteristic $\mathfrak{p}_k(\gamma) \cap \mathfrak{p}_k(\gamma)^{-1}$.

- Relationships among classes (e.g. $\text{PPol}_k \mathcal{S}^+$, $\text{PPol}_k \mathcal{S}$)
- Bases of identities
- Generating by a single structure
- Impact to level 2

- Relationships among classes (e.g. $\text{PPol}_k \mathcal{S}^+$, $\text{PPol}_k \mathcal{S}$)
- Bases of identities
- Generating by a single structure
- Impact to level 2

- Relationships among classes (e.g. $\text{PPol}_k \mathcal{S}^+$, $\text{PPol}_k \mathcal{S}$)
- Bases of identities
- Generating by a single structure
- Impact to level 2

- Relationships among classes (e.g. $\text{PPol}_k \mathcal{S}^+$, $\text{PPol}_k \mathcal{S}$)
- Bases of identities
- Generating by a single structure
- Impact to level 2

Let γ be a finite characteristic. We say that γ is **finitely determined** if there is a finite alphabet A such that for every finite alphabet B and all $u, v \in B^*$ we have:

$$((\forall \varphi : B \rightarrow A) \varphi(u) \gamma_A \varphi(v)) \text{ implies } u \gamma_B v .$$

Proposition

The following properties for a positive variety \mathcal{V} and the corresponding pseudovariety \mathbf{V} are equivalent.

- *\mathcal{V} is generated by a finite number of languages.*
- *\mathbf{V} is generated by a single ordered monoid.*
- *There exists a finite characteristic of \mathcal{V} which is finitely determined.*

Proposition

The positive variety $\text{PPol}_k \mathcal{S}^+$ is generated by a finite number of languages.

For $|A| = 2^{2k+1}$.

Proposition

The positive variety $\text{PPol}_1 \mathcal{S}$ is generated by a finite number of languages.

Proposition

The positive variety $\text{PPol}_2 \mathcal{S}$ is not generated by a finite number of languages.

Proposition

The positive variety $\text{PPol}_k \mathcal{S}^+$ is generated by a finite number of languages.

For $|A| = 2^{2k+1}$.

Proposition

The positive variety $\text{PPol}_1 \mathcal{S}$ is generated by a finite number of languages.

Proposition

The positive variety $\text{PPol}_2 \mathcal{S}$ is not generated by a finite number of languages.

Proposition

The positive variety $\text{PPol}_k \mathcal{S}^+$ is generated by a finite number of languages.

For $|A| = 2^{2k+1}$.

Proposition

The positive variety $\text{PPol}_1 \mathcal{S}$ is generated by a finite number of languages.

Proposition

The positive variety $\text{PPol}_2 \mathcal{S}$ is not generated by a finite number of languages.