UNIVERSITAT ROVIRA I VIRGILI

ROUGH APPROXIMATIONS IN VARIETIES OF REGULAR LANGUAGES

Gabriela Susana Martin Torres

Rough Approximations in Varieties of Regular Languages

Gabriela Martín Torres
Research Group on Mathematical Linguistics
Dep. de Filologies Romaniques. Universitat Rovira I Virgili.

November 26, 2015

Thesis for the degree of PhD in Formal Languages and Applications.

Thesis Supervisor: Magnus Steinby.

Thesis Co-Supervisor: María Dolores Jiménez López.

Universitat Rovira i Virgili

Rough Approximations in Varieties of Regular Languages

by

Gabriela Martín Torres

Submitted to the Research Group on Mathematical Linguistics
on July, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Formal Languages and Applications

Abstract

We study approximations of regular languages by members of a given variety $\mathcal{L}$ of regular languages. These are upper or lower approximations in the sense of Pawlak's rough set theory with respect to congruences belonging to the variety of congruences corresponding to $\mathcal{L}$. In particular, we consider the closest upper and lower approximations in $\mathcal{L}$. In so-called principal varieties these always exist, and we present algorithms for finding them, but for other varieties the situation is more complex. For non-principal $+$-varieties we study conditions for the existence of the closest upper and lower approximations. In particular, we consider varieties that are the union of a directed family of principal $+$-varieties, and pseudo-principal $+$-varieties, that are defined in this work.

Next, we consider the accuracy of the considered approximations, measured by the relative density of the object language in the approximation language and the asymptotic behavior of this quotient. In particular, we apply our measures of accuracy to k-definite, reverse k-definite, i, j-definite, k-testable and commutative approximations. Finally, we examine rough approximations under some infinite index indiscernibility relations as they were presented by Păun, Polkowski and Skowron (1997), looking at how they fit in our framework. We study their general features, comparing them with some of the families already studied, and in some cases introducing modifications in their definitions to make them congruences.

Although we consider mostly Eilenberg's $+$-varieties, the general ideas apply also to other types of varieties of languages. Our work may also be viewed as an approach to the characterizable inference problem in which a language of a certain kind is to be inferred from a given sample.

Thesis Supervisor: Magnus Steinby
Title: Professor Emeritus

Acknowledgments

When reviewing the persons around me whom I would like to acknowledge, I quickly realized how small and irrelevant this work really is compared to all I got from them during this endeavour.

My parents, Onofre y Mecha, who despite probably not understanding this words, were always there, giving me their all in each and every way as best they know.

Magnus Steinby, with his unshakeable kindness and patience, who showed me that being a great mathematician is just the smallest aspect of a great human being.

My good friends, Ivan, Daniela, Julia, who are much more than brothers to me, have been walking, deciphering life with me year after year for longer than I can remember. I cannot imagine the world without them.

Anna, my buddy, my sidekick, who taught me how to find joy and happiness every day, everywhere, no matter what. My small crazy adventures companion, my heart, my helmsman.

I want to dedicate this work to you all. It is nothing compared to all the love you have dedicated me.

Contents

Chapter 1

Introduction

In this work we consider the problem of approximating a given regular language by languages from a given family of regular languages. Here the target families will be $+$-varieties as defined by Eilenberg [12], i.e., varieties of regular languages that do not contain the empty word, but the general ideas apply equally well to other kinds of varieties of languages such as $*$-varieties [12] or varieties of regular tree languages (cf. [45, 46], for example). This kind of approximating languages may also be regarded as an approach to the inference of regular languages. The inference of a language from a sample is an important problem and many inference methods for various types of languages have been proposed (cf. [2, 15, 30], for example). Of special interest are the so-called characterizable inference methods that always produce a language belonging to a given family of languages. To infer a language in some $+$-variety $\mathcal{L}$ from a given sample S, we may either directly approximate S in $\mathcal{L}$, or then first obtain a regular extension R of S by some heuristic method, for example, and then approximate this R with a member of the variety $\mathcal{L}$.

As shown by Thérien [48, 49], for each $+$-variety $\mathcal{L}$ there is a corresponding variety of congruences $\mathcal{L}^c$ on the free semigroups generated by finite alphabets. A language belongs to $\mathcal{L}$ iff it is saturated by a congruence belonging to $\mathcal{L}^c$, and all the approximations that we consider are based on some congruence in $\mathcal{L}^c$. If R is a language over an alphabet X, then for any congruence θ on X^+, we define the lower θ-approximation R_θ and the upper θ-approximation R^θ of R as in Pawlak's rough set theory [34, 35] (cf. also [31] or [40]). Of special interest are the closest approximations, the greatest lower $\mathcal{L}$-approximation $R_{\mathcal{L}}$ and the least upper $\mathcal{L}$-approximation $R^{\mathcal{L}}$, and a part of our work is centered around them.

Some basic notions and our general notation is introduced in Chapter 2. In Chapter 3 we study and describe approximations of regular languages by members of several types of varieties $\mathcal{L}$ of regular languages. In 3.1 we start considering the upper and lower rough approximations R^{θ} and R_{θ} of a language R over an alphabet X with respect to a given congruence θ on the free semigroup X^{+}. These have naturally all the general properties of rough approximations. If θ is of finite index, they are always regular languages, and we present algorithms for finding them when R is a regular language given by a finite recognizer (and θ is effectively given).

In 3.2 we first recall Eilenberg's [12] $+$-varieties and Thérien's [49] corresponding varieties of congruences that we call simply $+$-filters. Then we define approximations in $+$-varieties. For any upper (lower) approximation of a given language R by a language L belonging to a $+$-variety $\mathcal{L}$, there is a congruence θ in the $+$-filter $\mathcal{L}^{c}$ that corresponds to $\mathcal{L}$ such that $R \subseteq R^{\theta} \subseteq L$ ($L \subseteq R_{\theta} \subseteq R$). In particular, if the least upper $\mathcal{L}$-approximation $R^{\mathcal{L}}$ (greatest lower $\mathcal{L}$-approximation $R_{\mathcal{L}}$) of R exists, then $R^{\mathcal{L}} = R^{\theta}$ ($R_{\mathcal{L}} = R_{\theta}$) for some congruence θ in $\mathcal{L}^{c}$.

In 3.3 we consider the case of principal varieties; a $+$-variety $\mathcal{L}$ is principal if the congruences in $\mathcal{L}^{c}$ corresponding to each alphabet form a principal filter. Since a smaller congruence always yields approximations that are at least as close to a given language as those given by a larger congruence, it follows that in a principal $+$-variety $\mathcal{L}$, the closest approximations $R^{\mathcal{L}}$ and $R_{\mathcal{L}}$ exist for every language R. We shall also present a few concrete examples of constructions of recognizers for approximations of a given language in a principal $+$-variety by using the algorithms of 3.1. For each such $+$-variety, a suitable practical formulation of the algorithm used is given.

For non-principal $+$-varieties the situation is far more complex. One of our propositions in 3.4 implies that in many well-known non-principal $+$-varieties $\mathcal{L}$, the closest approximations $R^{\mathcal{L}}$ or $R_{\mathcal{L}}$ exist only for languages R that themselves belong to $\mathcal{L}$. On the other hand, there are non-principal $+$-varieties such that a language may have one of the closest approximations while the other one does not exist.

In 3.5 we consider an important special case of non-principal +-varieties, the +-varieties that in Eilenberg's Variety Theorem [12] correspond to equationally defined varieties of finite semigroups. We show that these also can be viewed as natural generalizations of principal varieties. The main result is a theorem that tells when a given language has a least upper approximation in an equational +-variety. In 3.6 we present some concluding remarks of the chapter.

In Chapter 4 we study the accuracy of the considered approximations. In Section 4.1 we calculate the densities of the languages in the target +−varieties. In Section 4.2 we present two definitions of the accuracy of the upper rough approximation of a language. The first one is simply the quotient of the cardinality of the set of words up to a certain length in the language and the cardinality of the set of words up to the same length in the approximation language. The second is the limit of the first one when the length approaches infinity. Some basic properties and particular features of both definitions of the accuracy of approximations are shown. In Section 4.3 we look at the accuracy of approximations of languages over a one-letter alphabet and in Section 4.4 we establish the attainable accuracy values for approximations of languages of a given density over any alphabet, in the already mentioned target families. Finally, in Section 4.5, we make some concluding remarks.

Although the ideas of rough set theory have been applied in numerous areas, it seems that not much has been done along these lines in formal language theory. As a notable exception, we can mention the work by Păun, Polkowski and Skowron [32, 33]. These papers focus on approximating languages with respect some given similarity relations between words and the convergence of successive refinements of such approximations. In Chapter 5, we look into [33] in more detail. The relations defined by the authors are indiscernibility relations that are in some cases congruences, in some cases only tolerance relations, and they all have infinite index, because the related words are required to have the same length. We study some of them in Section 5.1, showing their relation to the families and rough approximations already considered. In Section 5.2, we present some modifications to make them fit better in the theoretical context of our work. Finally, in Section 5.3, we make some comments about the accuracy of the rough approximations shown in the chapter, comparing them with the accuracy of the rough approximations in the families they are closely related.

Chapter 2

Preliminaries

Sometimes we write $A := B$ to indicate that A is defined to be equal to B. The basic set-theoretic symbols $\cap$, $\cup$, $\subseteq$, $\subset$, $\ldots$ have their usual meanings. The complement $U \setminus A$ of a subset A of a given universe U is denoted by A', and the power set of A by $\wp(A)$. The cardinality of a set A is denoted $|A|$.

Let $\theta \subseteq A \times B$ be a relation from a set A to a set B. The fact that $(a,b) \in \theta$ for some $a \in A$ and $b \in B$, is expressed also by writing $a \, \theta \, b$. The *converse* of θ is the relation $\theta^{-1} := \{(b,a) \mid (a,b) \in \theta\}$. The *product* of θ and any $\rho \subseteq B \times C$ is the relation $\theta \circ \rho := \{(a,c) \mid (\exists b \in A) \, a \, \theta \, b, \; b \, \rho \, c\}$. The *diagonal relation* $\{(a,a) \mid a \in A\}$ and the *universal relation* $A \times A$ are denoted by Δ_A and ∇_A, respectively. A relation $\theta \subseteq A \times A$ is an *equivalence* on A if $\Delta_A \subseteq \theta$, $\theta^{-1} \subseteq \theta$ and $\theta \circ \theta \subseteq \theta$. Let $\mathrm{Eq}(A)$ be the set of all equivalences on A. For any $\theta \in \mathrm{Eq}(A)$, the *quotient set* A/θ is the set $\{[a]_\theta \mid a \in A\}$, where $[a]_\theta := \{b \in A \mid a \, \theta \, b\}$ is the θ-class of $a \in A$. The *natural mapping* $A \to A/\theta$, $a \mapsto [a]_\theta$ is denoted by ν_θ. If A/θ is finite, θ is said to be of *finite index.*

An equivalence $\theta \in \mathrm{Eq}(A)$ *saturates* a subset $H \subseteq A$ if H is the union of some θ-classes. Let $\mathrm{Sat}(\theta)$ denote the set of all subsets of A saturated by θ. The following facts are easy to prove.

2.0.1 Lemma *For any set A and any equivalences $\theta, \rho \in \mathrm{Eq}(A)$,*

(a) *if $\theta \subseteq \rho$, then* $\mathrm{Sat}(\theta) \supseteq \mathrm{Sat}(\rho)$, *and*

(b) $\mathrm{Sat}(\theta \vee \rho) = \mathrm{Sat}(\theta) \cap \mathrm{Sat}(\rho)$. □

For any mapping $\varphi\colon A \to B$, we often write $a\varphi$ for the image $\varphi(a)$ of an element $a \in A$. Furthermore, for any $H \subseteq A$ and any $K \subseteq B$, we set $H\varphi := \{a\varphi \mid a \in H\}$ and $K\varphi^{-1} := \{a \in A \mid a\varphi \in K\}$. If $\theta \in \mathrm{Eq}(B)$ is an equivalence on B, then $\varphi \circ \theta \circ \varphi^{-1}$ is the equivalence $\{(a_1, a_2) \in A \times A \mid a_1\varphi \, \theta \, a_2\varphi\}$ on A.

All the lattice theory needed here can be found in [7] or [10], for example. Partial orders are called simply *orders*. Hence an *ordered set* $(A, \leq)$ consists of a non-empty set A and a relation $\leq$ on A that is reflexive, antisymmetric and transitive. A *lattice* is an ordered set $(A, \leq)$ such that any two elements $a, b \in A$ have a least upper bound, the *join* $a \vee b$, and a greatest lower bound, the *meet* $a \wedge b$. A *complete lattice* is an ordered set $(A, \leq)$ such that the least upper bound $\sup H$ and the greatest lower bound $\inf H$ exist for every $H \subseteq A$. Recall that a *filter* of a lattice $(A, \leq)$ is a non-empty subset F of A such that (1) $a \leq b$ and $a \in F$ imply $b \in F$, and (2) $a \wedge b \in F$ whenever $a, b \in F$. The *filter generated* by a non-empty subset $H \subseteq A$, i.e., the least filter $[H)$ containing H, can easily be shown to be the set $\{a \in A \mid (\exists n > 0)(\exists b_1, \ldots b_n \in H)\, b_1 \wedge \ldots \wedge b_n \leq a\}$. The *principal filter* $[a) := \{x \in A \mid a \leq x\}$ generated by any given element $a \in A$ is the least filter that includes a as an element.

Let us now review some basic notions from the theory of finite automata and regular languages (cf. [1, 11, 12, 23, 22, 51], for example).

An *alphabet* is a non-empty set of symbols called *letters*. If X is an alphabet, then X^* denotes the set of all (finite) *words* over X, ε is the *empty word*, and X^+ is the set of non-empty words over X. Subsets of X^* are called *languages*, and subsets of X^+ are *ε-free* languages. As usual, X^* and X^+ stand, respectively, also for the free monoid and the free semigroup generated by X with concatenation as the operation. Unless stated otherwise, an alphabet is always assumed to be finite.

The length of a word $w \in X^*$ is denoted by $\lg(w)$, and $|w|_x$ is the number of appearances of a letter $x \in X$ in the word w. For any $k \geq 0$, let $X^k := \{w \in X^* \mid \lg(w) = k\}$, $X^{\geq k} := \{w \in X^* \mid \lg(w) \geq k\}$, $X^{\leq k} := \{w \in X^* \mid \lg(w) \leq k\}$, $X^{>k} := \{w \in X^* \mid \lg(w) > k\}$, and $X^{<k} := \{w \in X^* \mid \lg(w) < k\}$. For any integer

$k \geq 0$ and any word $w \in X^*$, we define the *k-prefix* $\mathrm{pref}_k(w)$ and the *k-suffix* $\mathrm{suff}_k(w)$ of w as follows:

(1) if $\mathrm{lg}(w) \leq k$, then $\mathrm{pref}_k(w) = \mathrm{suff}_k(w) = w$, and

(2) if $\mathrm{lg}(w) > k$, then $\mathrm{pref}_k(w)$ is the word u of length k such that $w = uv$ for some $v \in X^+$, and $\mathrm{suff}_k(w)$ is the word v of length k such that $w = uv$ for some $u \in X^+$.

Moreover, the set of *subwords* of w of length k is defined to be

$$\mathrm{sw}_k(w) := \{v \in X^k \mid (\exists u, u' \in X^*) w = uvu'\}.$$

Any alphabet X is also regarded as a set of unary operation symbols, and an *X-algebra* is then a system $\mathcal{A} = (A, X)$ in which A is a nonempty set and each letter $x \in X$ is interpreted as a unary operation $x^{\mathcal{A}} \colon A \to A$. The mappings $w^{\mathcal{A}} \colon A \to A$ induced by words $w \in X^*$ are obtained in the natural way: $\varepsilon^{\mathcal{A}}$ is the identity map 1_A on A, and $a(vx)^{\mathcal{A}} = (av^{\mathcal{A}})x^{\mathcal{A}}$ for any $a \in A$, $v \in X^*$ and $x \in X$. Any finite X-algebra $\mathcal{A} = (A, X)$ is also regarded as an *X-automaton*, and then A is its (finite) set of *states* and X is called its *input alphabet*. The *free X-algebra* $\mathcal{F}_X = (X^*, X)$ generated by $\{\varepsilon\}$ has the words over X as its elements and for each $x \in X$, the operation $x^{\mathcal{F}_X} \colon X^* \to X^*$ is defined by $ux^{\mathcal{F}_X} = ux$ ($u \in X^*$). Of course, $uv^{\mathcal{F}_X} = uv$ for all $u, v \in X^*$.

An equivalence $\theta \in \mathrm{Eq}(A)$ is a *congruence* on an X-algebra $\mathcal{A} = (A, X)$ if, for any $a, b \in A$, $a \,\theta\, b$ implies that $ax^{\mathcal{A}} \,\theta\, bx^{\mathcal{A}}$ for every $x \in X$. If θ is a congruence on $\mathcal{A}$, then the *quotient algebra* $\mathcal{A}/\theta = (A/\theta, X)$ defined by setting $[a]_\theta x^{\mathcal{A}/\theta} = [ax^{\mathcal{A}}]_\theta$, for any $a \in A$ and $x \in X$, is a well-defined X-algebra.

For any kind of algebra $\mathcal{A}$, let $\mathrm{Con}(\mathcal{A})$ and $\mathrm{FCon}(\mathcal{A})$ denote the sets of all congruences on $\mathcal{A}$ and all congruences on $\mathcal{A}$ of finite index, respectively. Obviously, $\mathrm{Con}(X^*) \subseteq \mathrm{Con}(\mathcal{F}_X)$ and $\mathrm{FCon}(X^*) \subseteq \mathrm{FCon}(\mathcal{F}_X)$ for any alphabet X. In fact, the congruences on the X-algebra $\mathcal{F}_X$ are exactly the right congruences on the monoid X^*.

An X-*recognizer* $\mathbf{A} = (\mathcal{A}, a_0, F)$ consists of an X-automaton $\mathcal{A} = (A, X)$, an *initial state* $a_0 \in A$, and a set of *final states* $F \subseteq A$. The *language recognized* by $\mathbf{A}$ is the set $L(\mathbf{A}) := \{w \in X^* \mid a_0 w^{\mathcal{A}} \in F\}$. A language $L \subseteq X^*$ is *recognizable*, or *regular*, if $L = L(\mathbf{A})$ for some X-recognizer $\mathbf{A}$. In this work we consider ε-free languages only. Let $Rec(X)$ denote the set of all ε-free regular languages over X, and let $Rec = \{Rec(X)\}_X$ be the family of all ε-free regular languages, where X ranges over all finite alphabets.

Chapter 3

Rough approximations in a +-variety

In this chapter, we study and describe approximations of regular languages by members of several types of varieties of regular languages. We start considering the upper and lower rough approximations R^θ and R_θ of a language R over an alphabet X with respect to a given congruence θ on the free semigroup X^+. As shown by Thérien [48, 49], for each $+$-variety $\mathcal{L}$ there is a corresponding variety of congruences $\mathcal{L}^c$ such that a language belongs to $\mathcal{L}$ iff it is saturated by a congruence belonging to $\mathcal{L}^c$, so that we consider approximations based on some congruence in $\mathcal{L}^c$. We then study $+$-varieties in general, for any upper (lower) approximation of a given language R by a language L belonging to a $+$-variety $\mathcal{L}$, there is a congruence θ in the $+$-filter $\mathcal{L}^c$ that corresponds to $\mathcal{L}$ such that $R \subseteq R^\theta \subseteq L$. Next we consider the case of principal varieties; a $+$-variety $\mathcal{L}$ is principal if the congruences in $\mathcal{L}^c$ corresponding to each alphabet form a principal filter, in this case the approximations always exist, and we show how to calculate them in some cases. For many well-known families of non-principal $+$-varieties, the closest approximations $R^{\mathcal{L}}$ or $R_{\mathcal{L}}$ exist only for languages R that themselves belong to $\mathcal{L}$. Finally, we study an important special case of non-principal $+$-varieties, the $+$-varieties that in Eilenberg's Variety Theorem [12] correspond to equationally defined varieties of finite semigroups. We present a theorem that tells when a given language has a least upper approximation in an equational $+$-variety. All the results, unless noted otherwise, come from our paper [27].

3.1 Rough approximations modulo a congruence

As it turns out that the approximations of languages that we want to find, actually are defined by certain congruences on the free semigroups X^+, we begin by introducing approximations of languages modulo such congruences. The following notions are derived from Pawlak's [34, 35] theory of *rough sets* (cf. also [31, 40]).

3.1.1 Definition Let θ be any equivalence relation on X^+ for some alphabet X. The *upper θ-approximation* of a language $R \subseteq X^+$ is the language $R^\theta := \bigcup\{[w]_\theta \mid w \in R\}$ and the *lower θ-approximation* of R is the language $R_\theta := \bigcup\{[w]_\theta \mid [w]_\theta \subseteq R\}$. □

The upper approximation R^θ is the union of all θ-classes that intersect with R, while R_θ is the union of all θ-classes totally contained in R. In the following lemma we list some well-known general properties of θ-approximations that do not depend on the nature of the universe of elements considered. Recall that $R' = X^+ \setminus R$ for any $R \subseteq X^+$.

3.1.2 Lemma *Let X be an alphabet and $\theta \in \mathrm{Eq}(X^+)$ be any equivalence on X^+. The following hold for any languages $L, R \subseteq X^+$.*

(a) $R_\theta, R^\theta \in \mathrm{Sat}(\theta)$ *and* $R_\theta \subseteq R \subseteq R^\theta$.

(b) $\emptyset_\theta = \emptyset = \emptyset^\theta$ *and* $(X^+)_\theta = X^+ = (X^+)^\theta$.

(c) $R_\theta = R$ *iff* $R^\theta = R$ *iff* $R \in \mathrm{Sat}(\theta)$.

(d) $(R_\theta)_\theta = (R_\theta)^\theta = R_\theta$ *and* $(R^\theta)_\theta = (R^\theta)^\theta = R^\theta$.

(e) $(L \cup R)_\theta \supseteq L_\theta \cup R_\theta$ *and* $(L \cup R)^\theta = L^\theta \cup R^\theta$.

(f) $(L \cap R)_\theta = L_\theta \cap R_\theta$ *and* $(L \cap R)^\theta \subseteq L^\theta \cap R^\theta$.

(g) *If $L \subseteq R$, then $L_\theta \subseteq R_\theta$ and $L^\theta \subseteq R^\theta$.*

(h) $(R')_\theta = (R^\theta)'$ *and* $(R')^\theta = (R_\theta)'$.

The inclusions in (e) *and* (f) *may be proper.* □

Assertions (e) and (f) of Lemma 3.1.2 hold more generally for any set $\mathcal{S} \subseteq \wp(X^+)$ of languages. In particular, $(\bigcup \mathcal{S})^\theta = \bigcup_{R\in\mathcal{S}} R^\theta$. This suggests that the upper θ-approximation of a language $R = \{w_1, w_2, w_3, \ldots\}$ can be obtained inductively as the limit of the sequence $(R_1)^\theta \subseteq (R_2)^\theta \subseteq (R_3)^\theta \subseteq \ldots$ of the upper θ-approximations of the finite subsets $R_n = \{w_1, \ldots, w_n\}$ $(n = 1, 2, 3, \ldots)$ of R. Indeed, $R^\theta = \bigcup_{n\geq 1}(R_n)^\theta$. On the other hand, the inclusion $\bigcup_{n\geq 1}(R_n)_\theta \subseteq R_\theta$ may be proper. In fact, it is proper always when R contains an infinite θ-class.

An equivalence θ on X^+ can also be given as the partition X^+/θ formed by the θ-classes, and every θ-approximation is the union of some θ-classes. The following simple lemma expresses the intuitively obvious fact that a finer partition yields closer approximations.

3.1.3 Lemma *Let $R \subseteq X^+$ for some alphabet X, and let $\theta, \varrho \in \mathrm{Eq}(X^+)$. If $\theta \subseteq \varrho$, then $R_\varrho \subseteq R_\theta \subseteq R \subseteq R^\theta \subseteq R^\varrho$. For $\theta \subset \varrho$, there exists a language $L \subseteq X^+$ such that $L^\theta \subset L^\varrho$, $L_\theta \supset L_\varrho$.*

Recall that the *syntactic congruence* σ_L of a language $L \subseteq X^+$ is defined by

$$u\,\sigma_L\,v \;\Leftrightarrow\; (\forall s, t \in X^*)(sut \in L \leftrightarrow svt \in L) \quad (u, v \in X^+),$$

and that it is the greatest congruence on X^+ that saturates L. Moreover, it is of finite index iff L is a regular language (cf. [12, 23, 38], for example). To prove the following basic fact it suffices to note that $\theta \subseteq \sigma_L$ for every $L \in \mathrm{Sat}(\theta)$.

3.1.4 Lemma *If $\theta \in \mathrm{FCon}(X^+)$, then $\mathrm{Sat}(\theta) \subseteq Rec(X)$. In particular, $R^\theta, R_\theta \in Rec(X)$ for any language $R \subseteq X^+$.* □

Any congruence θ on the semigroup X^+ can be extended to a congruence $\theta \cup \{(\varepsilon, \varepsilon)\}$ on the free X-algebra $\mathcal{F}_X = (X^*, X)$ by adding to it a new congruence class consisting of ε only. Let θ also denote this extended congruence. The quotient algebra $\mathcal{F}_X/\theta = (X^*/\theta, X)$ is defined by $[u]_\theta x^{\mathcal{F}_X/\theta} = [ux]_\theta$ $(u \in X^*, x \in X)$, and obviously

$$[\varepsilon]_\theta w^{\mathcal{F}_X/\theta} \;=\; [w]_\theta \quad \text{for every } w \in X^*.$$

This means that if $\theta \in \mathrm{FCon}(X^+)$, we get for any $L \in \mathrm{Sat}(\theta)$ an X-recognizer $\mathbf{F}_\theta(L) = (\mathcal{F}_X/\theta,\ [\varepsilon]_\theta,\ F(L))$ by letting $F(L)$ be $\{[w]_\theta \mid [w]_\theta \subseteq L\}$. In particular, the θ-approximations R^θ and R_θ of any language $R \subseteq X^+$ have recognizers of this kind. For R^θ the appropriate set of final states is $F(R^\theta) = \{[w]_\theta \mid w \in X^+,\ [w]_\theta \cap R \neq \emptyset\}$, and for R_θ it is $F(R_\theta) = \{[w]_\theta \mid w \in X^+,\ [w]_\theta \subseteq R\}$. Of course, these recognizers are not always minimal although they certainly are connected from the initial state.

Let us now find recognizers for R^θ and R_θ when $\theta \in \mathrm{FCon}(X^+)$ and $R \subseteq X^+$ is a regular language recognized by a given X-recognizer $\mathbf{A} = (\mathcal{A}, a_0, F)$, where $\mathcal{A} = (A, X)$. To determine the sets of final states $F(R^\theta), F(R_\theta) \subseteq X^*/\theta$, we consider the direct product $\mathcal{A} \times \mathcal{F}_X/\theta = (A \times X^*/\theta, X)$. Clearly,

$$(a_0, [\varepsilon]_\theta)w^{\mathcal{A}\times\mathcal{F}_X/\theta} \; = \; (a_0w^{\mathcal{A}}, [w]_\theta) \quad \text{for every } w \in X^*.$$

The sets $F(R_\theta)$ and $F(R^\theta)$ can now be found by computing the subalgebra

$$S \; := \; \{(a_0w^{\mathcal{A}}, [w]_\theta) \mid w \in X^*\}$$

of $\mathcal{A} \times \mathcal{F}_X/\theta$ generated by $(a_0, [\varepsilon]_\theta)$. Indeed,

$$F(R^\theta) \; = \; \{[v]_\theta \in X^*/\theta \mid (\exists a \in F)(a, [v]_\theta) \in S\}$$

and $F(R_\theta) \; = \; \{[v]_\theta \in X^*/\theta \mid (\forall a \in A)((a, [v]_\theta) \in S \to a \in F)\}$.

In the following procedure *ROUGH*, the set S is computed stepwise starting from the generator $(a_0, [\varepsilon]_\theta)$ and adding to it for any already found element $(a, [u]_\theta)$ of S (the loop starting at line (3)) and any $x \in X$ (the loop starting at line (4)) the next state $(ax^{\mathcal{A}}, [ux]_\theta)$ whenever it is a new element. The variable NEW holds all the elements of S that have not yet been used for extending S. When $(ax^{\mathcal{A}}, [ux]_\theta)$ has been formed for every $x \in X$ for a given $(a, [u]_\theta) \in NEW$, the pair $(a, [u]_\theta)$ is deleted from NEW, and it is never re-introduced there because it remains in S. In line (1), the variables are given their initial values. In particular, $F(R^\theta) := \emptyset$ and $F(R_\theta) := X^*/\theta - \{[\varepsilon]_\theta\}$ (note that $a_0 \notin F$). When a new pair $(b, [v]_\theta) \in S$ is found, we add $[v]_\theta$ to $F(R^\theta)$ if $b \in F$, but delete it from $F(R_\theta)$ if $b \notin F$. This is repeated until no new pairs are found.

3.1.5 Procedure *ROUGH*(**A**: X-recognizer, θ: congruence of finite index on X^+)
$\{\mathbf{A} = (\mathcal{A}, a_0, F)$, where $\mathcal{A} = (A, X)$, an X-recognizer for $R \subseteq X^+$; $\theta \in \mathrm{FCon}(X^+)$ (also extended to $\theta \in \mathrm{FCon}(\mathcal{F}_X))\}$

```
    var $S, NEW \subseteq A \times X^*/\theta$;   $F(R_\theta), F(R^\theta) \subseteq X^*/\theta$;   $a, b \in A$;   $x \in X$;   $u \in X^*$;
  (1) $S := \{(a_0, [\varepsilon]_\theta)\}$; $NEW := \{(a_0, [\varepsilon]_\theta)\}$; $F(R^\theta) := \emptyset$; $F(R_\theta) := X^+/\theta$;
  (2) while $NEW \neq \emptyset$ do begin
      (3) for $(a, [u]_\theta) \in NEW$ do begin
          (4) for $x \in X$ do begin
                  $b := ax^{\mathcal{A}}$;
                  if $(b, [ux]_\theta) \notin S$ then do begin
                  $S := S \cup \{(b, [ux]_\theta)\}$; $NEW := NEW \cup \{(b, [ux]_\theta)\}$;
              (5) if $b \in F$ then $F(R^\theta) := F(R^\theta) \cup \{[ux]_\theta\}$ else $F(R_\theta) := F(R_\theta) - \{[ux]_\theta\}$;
                  end {if} end {if}
              end {for}
              $NEW := NEW - \{(a, [u]_\theta)\}$;
          end {for}
      end {while}
return $(F(R_\theta), F(R^\theta))$
```

First of all, we note that for every $w \in X^*$ there is a word $v \in X^*$ such that $a_0 w^{\mathcal{A}} = a_0 v^{\mathcal{A}}$, $[w]_\theta = [v]_\theta$ and $(a_0 v^{\mathcal{A}}, [v]_\theta)$ is entered to S at some point. This can be shown by induction on the length of w. In other words, for every $w \in X^*$, the pair $(a_0 w^{\mathcal{A}}, [w]_\theta)$ is eventually entered to S, and since only pairs of this form are obtained, the procedure really computes the intended set S.

Let us now verify that in line (5) the right elements of X^*/θ are added to $F(R^\theta)$ or deleted from $F(R_\theta)$.

Consider any $[v]_\theta \in X^*/\theta - \{[\varepsilon]_\theta\}$. If $R \cap [v]_\theta \neq \emptyset$, then there is a word $w \in X^+$ such that $w \in R$ and $[w]_\theta = [v]_\theta$, and thus $(a_0, [\varepsilon]_\theta)w^{\mathcal{A}\times\mathcal{F}_X/\theta} = (b, [w]_\theta) = (b, [v]_\theta)$ for some $b \in F$. Now $(b, [v]_\theta) \in S$ and hence $[v]_\theta$ is added to $F(R^\theta)$ when the pair is encountered for the first time. On the other hand, if $R \cap [v]_\theta = \emptyset$, then $a_0 w^{\mathcal{A}} \notin F$ for every $w \in X^+$ such that $[w]_\theta = [v]_\theta$, and $[v]_\theta$ cannot be entered to $F(R^\theta)$. We may conclude that the value of $F(R^\theta)$ will be correct.

To show that also $F(R_\theta)$ finally gets the right value, we first assume that $[v]_\theta \subseteq R$. Then $a_0 w^{\mathcal{A}} \in F$ for every word such that $[w]_\theta = [v]_\theta$, and therefore $[v]_\theta$ is never deleted from $F(R_\theta)$. On the other hand, if $[v]_\theta \nsubseteq R$, then there is a word $w \in X^+$ such that $w \notin R$ and $[w]_\theta = [v]_\theta$, and then $(a_0, [\varepsilon]_\theta)w^{\mathcal{A}\times\mathcal{F}_X/\theta} = (b, [v]_\theta)$ for some $b \notin F$. Therefore $[v]_\theta$ is correctly deleted from $F(R_\theta)$ when the pair $(b, [v]_\theta)$ is formed for the first time.

If $|A| = n$, $|X^*/\theta| = m$ and $|X| = k$, then the inner **for**-loop is iterated at most nmk times. The dominating term in the time estimate for each iteration is the time needed for computing $[ux]_\theta$ from $[u]_\theta$ and x, and this depends naturally much on the congruence θ and how it is given.

The following alternative method for computing θ-approximations uses the inverse transition function of the recognizer $\mathbf{A} = (\mathcal{A}, a_0, F)$ of the given language R. Only the upper approximation is computed but by Lemma 3.1.2 (h) the lower approximation can be obtained by the same method. Again, one constructs the set S of pairs of the form $(a_0 w^{\mathcal{A}}, [w]_\theta)$, but now some redundancy can be avoided by tracing computations of $\mathbf{A}$ backwards starting from the pairs $(a, [\varepsilon]_\theta)$ with $a \in F$, and thus forming words w for which $a_0 w^{\mathcal{A}} \in F$ backwards by extending their suffixes letter by letter.

3.1.6 Procedure *IROUGH*($\mathbf{A}$: X-recognizer, θ: congruence of finite index on X^+) $\{\mathbf{A} = (\mathcal{A}, a_0, F)$, where $\mathcal{A} = (A, X)$, an X-recognizer for $R \subseteq X^+$; $Inv\colon A \times X \to \wp(A)$ the inverse transition function of $\mathbf{A}$; $\theta \in \mathrm{FCon}(X^+)$ (extended to $\theta \in \mathrm{FCon}(\mathcal{F}_X)\}$

```
    **var** $S, NEW \subseteq A \times X^*/\theta;\quad F(R^\theta) \subseteq X^*/\theta;\quad a, b \in A;\; x \in X;\quad u \in X^*;$

  (1) **for** $b \in A$ **do for** $x \in X$ **do** $Inv(b, x) := \{a \in A \mid ax^{\mathcal{A}} = b\};$

  (2) $S := \{(a, [\varepsilon]_\theta) \mid a \in F\};\; NEW := S;\; F(R^\theta) := \emptyset;$

  (3) **while** $NEW \neq \emptyset$ **do begin**

      (3) **for** $(a, [u]_\theta) \in NEW$ **do begin**

          (4) **for** $x \in X$ **do begin**

              (5) **for** $b \in Inv(a, x)$ **and** $(b, [xu]_\theta) \notin S$ **do begin**

                  $S := S \cup \{(b, [xu]_\theta)\};\; NEW := NEW \cup \{(b, [xu]_\theta)\};$

                  **if** $b = a_0$ **then** $F(R^\theta) := F(R^\theta) \cup \{[xu]_\theta\}$

                  **end** {for}

              **end** {for}

              $NEW := NEW - \{(a, [u]_\theta)\};$

          **end** {for}

      **end** {while}

**return** $(F(R^\theta))$
```

Let us now verify that the subset $F(R^\theta)$ is correctly constructed. For this we consider any $[v]_\theta \in X^*/\theta - \{[\varepsilon]_\theta\}$.

If $R \cap [v]_\theta \neq \emptyset$, then there is a word $w \in X^+$ such that $w \in R$ and $[w]_\theta = [v]_\theta$. Let $w = x_0x_1 \ldots x_n$ for some $n \geq 0$ and $x_0, x_1, \ldots, x_n \in X$, and let $a_0, a_1, \ldots, a_{n+1}$ be the sequence of states that **A** assumes when accepting w. Since $a_{n+1} \in F$, the pair $(a_{n+1}, [\varepsilon]_\theta)$ is entered into the initial sets S and NEW. Moreover, $a_{k-1} \in Inv(a_k, x_{k-1})$ for every $k = n+1, n, \ldots, 1$. Hence, we can show by induction on k that $(a_{k-1}, [u_k]_\theta)$, where $u_k = x_{k-1} \ldots x_n$, is entered to NEW, unless it is already in S, for every $k = n+1, n, \ldots, 1$. Since $u_1 = w$, we will eventually get $(a_0, [w]_\theta) \in NEW$, and hence $[v]_\theta (= [w]_\theta)$ is added to $F(R^\theta)$.

Let us now suppose that $[v]_\theta \in F(R^\theta)$, that is to say, $(a_0, [v]_\theta)$ was added to NEW at some step. This means that there is a word $w = x_0 x_1 \ldots x_n$ such that $[w]_\theta = [v]_\theta$ and, if $a_0, a_1, \ldots, a_{n+1}$ is the sequence of states that $\mathbf{A}$ assumes when reading w, then $a_{n+1} \in F$ and $a_{k-1} \in Inv(a_k, x_{k-1})$ for every $k = n+1, n, \ldots, 1$. But this means that $a_0 w^{\mathcal{A}} \in F$ and $w \in R$, and therefore $R \cap [v]_\theta \neq \emptyset$.

3.2 Rough approximations in a $+$-variety

In this section we introduce the problem of approximating a regular language by a language belonging to a given variety of regular languages. First we review the parts of Eilenberg's variety theory needed here fixing at the same time our terminology and notation. Full expositions can be found in [1, 12, 23, 38, 39], for example.

A *$+$-variety* $\mathcal{L} = \{\mathcal{L}(X)\}_X$ assigns to each alphabet X a non-empty set $\mathcal{L}(X) \subseteq Rec(X)$ of ε-free regular languages over X in such a way that for all X and Y,

(1) $L \cap R, L' \in \mathcal{L}(X)$ whenever $L, R \in \mathcal{L}(X)$,

(2) $L \in \mathcal{L}(X)$ implies that the quotient languages $w^{-1}L := \{u \in X^+ \mid wu \in L\}$ and $Lw^{-1} := \{u \in X^+ \mid uw \in L\}$ are also in $\mathcal{L}(X)$ for every $w \in X^+$, and

(3) $L \in \mathcal{L}(Y)$ implies $L\varphi^{-1} \in \mathcal{L}(X)$ for every homomorphism $\varphi : X^+ \to Y^+$.

The homomorphisms in (3) never shorten a word. Note also that we excluded the possibility that $\mathcal{L}(X) = \emptyset$ for some X. Hence, the least $+$-variety is $Triv = \{Triv(X)\}_X$, where $Triv(X) := \{\emptyset, X^+\}$ for each X. It is obvious that the class $\mathbf{VRL}^+$ of all $+$-varieties forms a complete lattice $(\mathbf{VRL}^+, \subseteq)$ when inclusion is defined by the natural alphabetwise condition: $\mathcal{K} \subseteq \mathcal{L}$ iff $\mathcal{K}(X) \subseteq \mathcal{L}(X)$ for every X.

Eilenberg's fundamental Variety Theorem [12] establishes a bijection between $+$-varieties and varieties of finite semigroups (pseudovarieties) thus describing the families of regular ε-free languages that can be characterized by syntactic semigroups. However, we shall use the following description of $+$-varieties by means of certain systems of congruences added to the theory by Thérien [48, 49].

A *+-variety of filters of congruences*, called here a *+-filter* for short, $\Gamma = \{\Gamma(X)\}_X$ assigns to each alphabet X a nonempty set $\Gamma(X) \subseteq \mathrm{FCon}(X^+)$ of congruences of finite index on X^+ in such a way that for all alphabets X and Y,

(1) $\Gamma(X)$ is a filter of $\mathrm{FCon}(X^+)$, and

(2) if $\varphi\colon X^+ \to Y^+$ is a homomorphism and $\theta \in \Gamma(Y)$, then $\varphi \circ \theta \circ \varphi^{-1} \in \Gamma(X)$.

If $\mathbf{VFC}^+$ denotes the class of all +-filters, then $(\mathbf{VFC}^+, \subseteq)$ is a complete lattice for the natural alphabetwise defined $\subseteq$-relation.

The mappings connecting $\mathbf{VRL}^+$ and $\mathbf{VFC}^+$ are defined as follows. For any +-variety $\mathcal{L} = \{\mathcal{L}(X)\}_X$ and any +-filter $\Gamma = \{\Gamma(X)\}_X$, let

(1) $\mathcal{L}^c$ be the +-filter such that for each X, $\mathcal{L}^c(X) := [\{\sigma_L \mid L \in \mathcal{L}(X)\})$ is the filter of $\mathrm{FCon}(X^+)$ generated by the syntactic congruences of the members of $\mathcal{L}(X)$, and

(2) Γ^l be the +-variety where $\Gamma^l(X) := \{L \subseteq X^+ \mid \sigma_L \in \Gamma(X)\}$ for each X.

If we omit varieties of finite semigroups, the Variety Theorem reads as follows.

3.2.1 Proposition *The mappings $\mathcal{L} \mapsto \mathcal{L}^c$ and $\Gamma \mapsto \Gamma^l$ form a pair of mutually inverse isomorphisms between the lattices $(\mathbf{VRL}^+, \subseteq)$ and $(\mathbf{VFC}^+, \subseteq)$. That is to say, both maps are order-preserving, and*

(a) *$\mathcal{L}^c \in \mathbf{VFC}^+$ and $\mathcal{L}^{cl} = \mathcal{L}$ for every +-variety $\mathcal{L}$, and*

(b) *$\Gamma^l \in \mathbf{VRL}^+$ and $\Gamma^{lc} = \Gamma$ for every +-filter Γ.* □

Let us also recall the following facts that explain why many +-varieties are most naturally defined in terms of the corresponding +-filters.

3.2.2 Lemma *Let $\mathcal{L}$ be a +-variety. For any X and $L \subseteq X^+$,*

(a) *$L \in \mathcal{L}(X)$ iff $\sigma_L \in \mathcal{L}^c(X)$, and*

(b) *$L \in \mathcal{L}(X)$ iff $L \in \mathrm{Sat}(\theta)$ for some $\theta \in \mathcal{L}^c(X)$.* □

Furthermore, we also have the following fact.

3.2.3 Lemma *Let $\mathcal{L}$ be a $+$-variety and X be any alphabet. For any $\theta \in \mathrm{FCon}(X^+)$, $\mathrm{Sat}(\theta) \subseteq \mathcal{L}(X)$ iff $\theta \in \mathcal{L}^c(X)$.*

Proof. If $\mathrm{Sat}(\theta) \subseteq \mathcal{L}(X)$, then $[w]_\theta \in \mathcal{L}(X)$, and hence $\sigma_{[w]_\theta} \in \mathcal{L}^c(X)$, for every $w \in X^+$. Therefore also $\theta \in \mathcal{L}^c(X)$ because $\theta = \bigcap\{\sigma_{[w]_\theta} \mid w \in X^+\}$ and the number of different θ-classes $[w]_\theta$ is finite. Assume then that $\theta \in \mathcal{L}^c(X)$. If $L \in \mathrm{Sat}(\theta)$, then $\theta \subseteq \sigma_L$, and hence $L \in \mathcal{L}(X)$. □

We shall now introduce the approximations that we are mainly interested in. In the rest of this section, $\mathcal{L} = \{\mathcal{L}(X)\}_X$ is always a fixed, but arbitrarily chosen, $+$-variety, and X and Y are any alphabets.

3.2.4 Definition A language L is called an *upper $\mathcal{L}$-approximation* of a language $R \subseteq X^+$ if $R \subseteq L \in \mathcal{L}(X)$. The *least upper $\mathcal{L}$-approximation* of R is an upper $\mathcal{L}$-approximation L of R such that $L \subseteq K$ for every upper $\mathcal{L}$-approximation K of R, and when it exists, it is denoted by $R^{\mathcal{L}}$. The *lower $\mathcal{L}$-approximations* and the *greatest lower $\mathcal{L}$-approximation* $R_{\mathcal{L}}$ of R are defined dually. □

Clearly, a language R has at most one least upper $\mathcal{L}$-approximation and at most one greatest lower $\mathcal{L}$-approximation, and hence the symbols $R^{\mathcal{L}}$ and $R_{\mathcal{L}}$ are justified.

A *minimal upper $\mathcal{L}$-approximation* of R is naturally an upper $\mathcal{L}$-approximation L of R for which there is no $K \in \mathcal{L}(X)$ such that $R \subseteq K \subset L$, and *maximal lower $\mathcal{L}$-approximations* are defined correspondingly. However, the following lemma shows that these notions are of no use here.

3.2.5 Lemma *If a language $R \subseteq X^+$ has a minimal upper $\mathcal{L}$-approximation, this is also the least upper $\mathcal{L}$-approximation of R. Similarly, if R has a maximal lower $\mathcal{L}$-approximation, it is the greatest lower $\mathcal{L}$-approximation of R.*

Proof. Assume that R has a minimal upper $\mathcal{L}$-approximation L that is not $R^{\mathcal{L}}$. Then there is an upper $\mathcal{L}$-approximation K of R such that $L \subseteq K$ does not hold. However, $L \cap K \in \mathcal{L}(X)$ since $\mathcal{L}$ is a $+$-variety, and hence $L \cap K$ is an upper $\mathcal{L}$-approximation of

R properly included in L, a contradiction. The assertion about lower approximations follows similarly from the fact that $K \cup L \in \mathcal{L}(X)$ for any $K, L \in \mathcal{L}(X)$. □

The following proposition shows that for any $+$-variety $\mathcal{L}$, all $\mathcal{L}$-approximations are determined by the congruences in $\mathcal{L}^c$.

3.2.6 Proposition *For any languages $R, L \subseteq X^+$,*

(a) *L is an upper $\mathcal{L}$-approximation of R iff $R \subseteq L$ and $L \in \mathrm{Sat}(\theta)$ for some $\theta \in \mathcal{L}^c(X)$;*

(b) *L is a lower $\mathcal{L}$-approximation of R iff $L \subseteq R$ and $L \in \mathrm{Sat}(\theta)$ for some $\theta \in \mathcal{L}^c(X)$.*

Proof. If L is an upper $\mathcal{L}$-approximation of R, then $R \subseteq L$ and $L \in \mathcal{L}(X)$, and hence $\sigma_L \in \mathcal{L}^c(X)$. Since $L \in \mathrm{Sat}(\sigma_L)$, this proves one direction of (a). On the other hand, if $R \subseteq L$ and $L \in \mathrm{Sat}(\theta)$ for some $\theta \in \mathcal{L}^c(X)$, then $L \in \mathcal{L}(X)$ by Lemma 3.2.2, and hence L is an upper $\mathcal{L}$-approximation of R. Statement (b) has a similar proof. □

3.2.7 Corollary *For any $R \subseteq X^+$ and any $\theta \in \mathcal{L}^c(X)$, R^θ is an upper $\mathcal{L}$-approximation of R and R_θ is a lower $\mathcal{L}$-approximation of R.* □

Next we show that every $\mathcal{L}$-approximation either is of the above kind or then it can be replaced with a closer approximation of this type.

3.2.8 Proposition *Consider any language $R \subseteq X^+$. If K is any lower $\mathcal{L}$-approximatio of R and L is any upper $\mathcal{L}$-approximation of R, then there is a congruence $\theta \in \mathcal{L}^c(X)$ such that $K \subseteq R_\theta \subseteq R \subseteq R^\theta \subseteq L$.*

Proof. Let $\theta := \sigma_K \cap \sigma_L$. Then $\theta \in \mathcal{L}^c(X)$ because $\sigma_K, \sigma_L \in \mathcal{L}^c(X)$. Moreover,

$$K = K_{\sigma_K} \subseteq R_{\sigma_K} \subseteq R_\theta \subseteq R \subseteq R^\theta \subseteq R^{\sigma_L} \subseteq L^{\sigma_L} = L$$

by Lemma 3.1.2(g) and Lemma 3.1.3 because $\theta \subseteq \sigma_K$, $\theta \subseteq \sigma_L$ and $K \subseteq R \subseteq L$. □

3.2.9 Corollary *Let $R \subseteq X^+$ for some X. If $R^{\mathcal{L}}$ exists, then $R^{\mathcal{L}} = R^{\theta}$ for some $\theta \in \mathcal{L}^c(X)$. Similarly, if $R_{\mathcal{L}}$ exists, then $R_{\mathcal{L}} = R_{\theta}$ for some $\theta \in \mathcal{L}^c(X)$. Moreover, if both $R^{\mathcal{L}}$ and $R_{\mathcal{L}}$ exist, then $R^{\mathcal{L}} = R^{\theta}$ and $R_{\mathcal{L}} = R_{\theta}$ for some $\theta \in \mathcal{L}^c(X)$.* □

Next we note a connection between least upper $\mathcal{L}$-approximations and greatest lower $\mathcal{L}$-approximations.

3.2.10 Proposition *Let $R \subseteq X^+$ for some X. Then $R^{\mathcal{L}}$ exists iff $(R')_{\mathcal{L}}$ exists, and then $R^{\mathcal{L}} = ((R')_{\mathcal{L}})'$. Similarly, $R_{\mathcal{L}}$ exists iff $(R')^{\mathcal{L}}$ exists, and then $R_{\mathcal{L}} = ((R')^{\mathcal{L}})'$.*

Proof. If $R^{\mathcal{L}}$ exists, then $R^{\mathcal{L}} = R^{\theta}$ for some $\theta \in \mathcal{L}^c(X)$. We claim that $(R^{\theta})' = (R')_{\mathcal{L}}$. Of course, $(R^{\theta})' = (R')_{\theta}$ is a lower $\mathcal{L}$-approximation of R'. If $(R^{\theta})'$ were not the greatest lower $\mathcal{L}$-approximation of R', then there would exist a $\rho \in \mathcal{L}^c(X)$ such that $(R^{\theta})' \subset (R')_{\rho} \subseteq R'$. However, this would imply that $R^{\theta} \supset ((R')_{\rho})' = R^{\rho} \supseteq R$, contradicting the assumption that $R^{\theta} = R^{\mathcal{L}}$.

Assume now that $(R')_{\mathcal{L}}$ exists. Then $(R')_{\mathcal{L}} = (R')_{\theta}$ for some $\theta \in \mathcal{L}^c(X)$, and it can be seen that R^{θ} is $R^{\mathcal{L}}$. The second assertion has a similar proof. □

3.2.11 Corollary *The least upper $\mathcal{L}$-approximation $R^{\mathcal{L}}$ exists for every (regular) language $R \subseteq X^+$ iff the greatest lower $\mathcal{L}$-approximation $R_{\mathcal{L}}$ exists for every (regular) language $R \subseteq X^+$.* □

3.3 Approximations in principal $+$-varieties

We shall now consider $+$-varieties of a special kind, the so-called principal $+$-varieties. There are many examples of these and many further $+$-varieties are naturally given as unions of principal $+$-varieties. For more about principal varieties, cf. [37, 45, 46].

A $+$-filter Γ is called *principal*, if for each alphabet X, $\Gamma(X)$ is a principal filter, i.e., $\Gamma(X) = [\gamma(X))$ for some congruence $\gamma(X) \in \mathrm{FCon}(X^+)$. A $+$-variety $\mathcal{L}$ is called *principal* if $\mathcal{L}^c$ is a principal $+$-filter. The following lemma (cf. [45]) can be used for verifying that a system of congruences yields a principal $+$-filter.

3.3.1 Lemma *Assume that we are given a congruence $\gamma(X) \in \mathrm{FCon}(X^+)$ for each alphabet X. Then $\Gamma = \{[\gamma(X))\}_X$ is a principal $+$-filter iff $\gamma(X) \subseteq \varphi \circ \gamma(Y) \circ \varphi^{-1}$ for all X and Y and every homomorphism $\varphi : X^+ \to Y^+$.*

By applying the above condition to the endomorphisms $\varphi : X^+ \to X^+$, we see that in a principal $+$-filter $\Gamma = \{[\gamma(X))\}_X$, the congruences $\gamma(X)$ are fully invariant. The following basic fact is an immediate consequence of Lemma 3.1.3 and Proposition 3.2.8.

3.3.2 Proposition *For any principal $+$-variety $\mathcal{L}$, all least upper $\mathcal{L}$-approximations and all greatest lower $\mathcal{L}$-approximations exist. More precisely: if $\mathcal{L}^c = \{[\gamma(X))\}_X$, then $R^{\mathcal{L}} = R^{\gamma(X)}$ and $R_{\mathcal{L}} = R_{\gamma(X)}$ for any X and $R \subseteq X^+$.*

In the next examples, we consider two well-known families of principal $+$-varieties.

3.3.3 Example For any $k \geq 0$ and any X, we define the relation $\delta_k(X)$ on X^+ by

$$u \, \delta_k(X) \, v \;\Leftrightarrow\; \mathrm{suff}_k(u) = \mathrm{suff}_k(v) \quad (u, v \in X^+).$$

Clearly, $\delta_k(X) \in \mathrm{FCon}(X^+)$ for every $k \geq 0$, and

$$X^+/\delta_k(X) \;=\; \{\{w\} \mid w \in X^{<k}, w \neq \varepsilon\} \cup \{X^*w \mid w \in X^k\},$$

i.e., each word $w \in X^+$ of length $< k$ forms a singleton class $\{w\}$ and each word w of length k determines the class X^*w of all words ending in w. In particular, $X^+/\delta_0(X) = \{X^+\}$. Moreover, by using Lemma 3.3.1 it is easy to see that $kDef^c := \{[\delta_k(X))\}_X$ is a principal $+$-filter. A language $L \subseteq X^+$ is *k-definite* [17, 36] if it is saturated by $\delta_k(X)$, that is to say, if the membership of a word w in L is determined by $\mathrm{suff}_k(w)$. Hence, the family of k-definite (ε-free) languages $kDef = \{kDef(X)\}_X$ is the principal $+$-variety corresponding to the $+$-filter $kDef^c$. Because $\delta_k(X) \supset \delta_{k+1}(X)$ for every $k \geq 0$ and any X, we have a properly ascending chain $0Def^c \subset 1Def^c \subset 2Def^c \subset \ldots$ of principal $+$-filters and a corresponding chain $0Def \subset 1Def \subset 2Def \subset \ldots$ of principal $+$-varieties. For any language $R \subseteq X^+$ and each $k \geq 0$,

$$R^{kDef} \;=\; R^{\delta_k(X)} \;=\; (R \cap X^{<k}) \cup \bigcup \{X^* w \mid w \in X^k, R \cap X^* w \neq \emptyset\},$$

and similarly

$$R_{kDef} \;=\; R_{\delta_k(X)} \;=\; (R \cap X^{<k}) \cup \bigcup \{X^* w \mid w \in X^k, X^* w \subseteq R\}.$$

In particular, $R^{0Def} = X^+$ for any $R \neq \emptyset$, $R_{0Def} = \emptyset$ for any $R \subset X^+$ and $\emptyset^{0Def} = \emptyset$, $X^+_{0Def} = X^+$. □

3.3.4 Example The following definitions, notation and results were taken from [24] (cf. also [42], [12] (VIII.9) and [38], Chapter 4.1).

We say that a word $w = x_1 ... x_n \in X^*$, where $x_1, ..., x_n \in X$, is a *scattered subword* of $u \in X^*$ if $u = u_1 x_1 u_2 x_2 ... x_n u_{n+1}$ for some $u_1, ..., u_{n+1} \in X^*$. The set of scattered subwords of length (at most) k of a word u is denoted by $\mathrm{ssw}_k(u)$ ($\mathrm{ssw}_{\leqslant k}(u)$). Let $J_k(X)$ be the relation defined on X^+ by

$$(u, v) \in J_k(X) \Leftrightarrow \mathrm{ssw}_{\leqslant k}(u) = \mathrm{ssw}_{\leqslant k}(v)$$

Observe that if the length of a word u is greater than k and has the same set of scattered subwords of length k as another word v, then $\mathrm{ssw}_{\leqslant k}(u) = \mathrm{ssw}_{\leqslant k}(v)$, thus the above definition splits into two parts:

$\forall u, v \in X^+$, if $\lg(u) < k$, then $(u, v) \in J_k(X)$ iff $u = v$.

$\forall u, v \in X^+$, if $\lg(u) \geqslant k$, then $(u, v) \in J_k(X)$ iff $\mathrm{ssw}_k(u) = \mathrm{ssw}_k(v)$.

The *shuffle* of $u, v \in X^*$ is the set $u \circ v = \{u_1 v_1 ... u_n v_n \mid u_1, ..., u_n, v_1, ..., v_n \in X^*, u = u_1 ... u_n, v = v_1 ... v_n, n \geqslant 0\}$. The *shuffle* of two languages $A, B \subseteq X^*$ is $A \circ B = \bigcup_{u \in A, v \in B} u \circ v$.

For any alphabet X and $k \geqslant 0$, $J_k(X) \in \mathrm{FCon}(X^+)$, and

$$X^+/J_k(X) = \{\{u\} \mid u \in X^{<k}\} \cup \{(\bigcap_{s\in S} s \circ X^* \setminus \bigcup_{t\in X^k\setminus S} t \circ X^*) \mid S \subseteq X^k\}$$

Then, each word u of length $< k$ forms a singleton class, and each word u of length $\geqslant k$, determines the class $\bigcap_{s\in S} s \circ X^* \setminus \bigcup_{t\in X^k\setminus S} t \circ X^*$, where $S = \mathrm{ssw}_k(u)$.

For any morphism $\phi : X^+ \to Y^+$, if $(u, v) \in J_k(X)$ then $(u\phi, v\phi) \in J_k(Y)$, because $\mathrm{ssw}_k(u) = \mathrm{ssw}_k(v)$ obviously implies $\mathrm{ssw}_k(u\phi) = \mathrm{ssw}_k(v\phi)$. Applying lemma 3.3.1, we can conclude that the family $\mathcal{J}_k^c = \{[J_k(X))\}_X$ is a principal +-filter.

A language $R \subseteq X^+$ is *piecewise k-testable* if it is saturated by $J_k(X)$, that is to say, if the membership of a word u in R is determined by $\mathrm{ssw}_k(u)$.

Hence, the family of piecewise k-testable (ε-free) languages $\mathcal{J}_k = \{\mathcal{J}_k(X)\}_X$ is the principal +-variety corresponding to the principal +-filter $\mathcal{J}_k^c$ and thus, for every regular language $R \subseteq X^+$, $k \geqslant 0$, the least upper $\mathcal{J}_k$-approximation and the greatest lower $\mathcal{J}_k$-approximation exist. Moreover,

$$R^{\mathcal{J}_k} = R^{J_k(X)} = (R \cap X^{<k}) \cup \{(\bigcap_{s\in \mathrm{ssw}_k(u)} s \circ X^* \setminus \bigcup_{t\in X^k\setminus \mathrm{ssw}(u)} t \circ X^*) \mid u \in R \cap X^{\geqslant k}\}$$

and

$$R_{\mathcal{J}_k} = R_{J_k(X)} =$$

$$(R \cap X^{<k}) \cup \{(\bigcap_{s\in S} s \circ X^* \setminus \bigcup_{t\in X^k\setminus S} t \circ X^*) \mid S \subseteq X^k, (\bigcap_{s\in S} s \circ X^* \setminus \bigcup_{t\in X^k\setminus S} t \circ X^*) \subseteq R\}.$$

A language is *piecewise testable* if it is piecewise k-testable for some $k \geqslant 0$. As $J_k(X) \supset J_{k+1}(X)$ for every $k \geq 0$ and any X, we have a properly ascending chain $\mathcal{J}_0^c \subset \mathcal{J}_1^c \subset \mathcal{J}_2^c \subset \ldots$ of principal +-filters and a corresponding chain $\mathcal{J}_0 \subset \mathcal{J}_1 \subset \mathcal{J}_2 \subset \ldots$ of principal +-varieties. Hence, the union $\mathcal{J} = \bigcup_{k\geqslant 1}\{\mathcal{J}_k(X)\}_X$ is a non-principal +-variety. We consider approximations in this type of variety in the next section.

Let $\mathcal{L}$ be a principal $+$-variety defined by a principal $+$-filter $\mathcal{L}^c = \{[\gamma(X))\}_X$ and let $R \subseteq X^+$ be a regular language. By Proposition 3.3.2, recognizers for $R^{\mathcal{L}}$ and $R_{\mathcal{L}}$ can be constructed from the quotient algebra $\mathcal{F}_X/\gamma(X)$ as described in Section 3.1 by applying *ROUGH* or *IROUGH* to the congruence $\gamma(X)$. When using *ROUGH* in this situation, it may be convenient to replace the quotient algebra $\mathcal{F}_X/\gamma(X)$ with an isomorphic X-algebra $\mathcal{F}(\gamma(X)) = (W(\gamma(X)), X)$ in such a way that the isomorphism is given by a map $[w]_{\gamma(X)} \mapsto \widehat{w}$ assigning to each $\gamma(X)$-class $[w]_{\gamma(X)}$ an element $\widehat{w}$ of $W(\gamma(X))$ that in a natural way identifies the class. Moreover, it has to be assumed that there is an effective procedure to compute the representative $\widehat{w} \in W(\gamma(X))$ of $[w]_{\gamma(X)}$ for any given word $w \in X^*$. When *IROUGH* is used, we don't make use of the algebraic structure of $\mathcal{F}_X/\gamma(X)$, and hence it suffices to introduce a suitable set $W(\gamma(X))$ of representatives for the $\gamma(X)$-classes. Of course, the sets $F(R^{\gamma(X)})$ and $F(R_{\gamma(X)})$ are now subsets of $W(\gamma(X))$.

For example, if we want to find the least upper k-definite approximation of a language $R \subseteq X^+$, the congruence to consider is $\delta_k(X)$, and we can take $W(\delta_k(X))$ to be $X^{\leq k}$ with $\widehat{w} = \mathrm{suff}_k(w)$ for each $w \in X^*$, and the algebra $\mathcal{F}(\delta_k(X)) = (W(\delta_k(X)), X)$ is defined by setting $wx^{\mathcal{F}(\delta_k(X))} = \mathrm{suff}_k(wx)$ for any $w \in X^{\leq k}$ and $x \in X$.

3.3.5 Example Let us find the least upper 2-definite approximation of the language $R = 01^*0$ over the alphabet $X = \{0, 1\}$. It is recognized by the X-recognizer $\mathbf{A} = (\mathcal{A}, a_0, F)$ with $A = \{a_0, a_1, a_{tr}, a_f\}$, $F = \{a_f\}$ and the transitions defined by $a_0 0^{\mathcal{A}} = a_1$, $a_0 1^{\mathcal{A}} = a_{tr}$, $a_1 0^{\mathcal{A}} = a_f$, $a_1 1^{\mathcal{A}} = a_1$, $a_{tr} 0^{\mathcal{A}} = a_{tr} 1^{\mathcal{A}} = a_f 0^{\mathcal{A}} = a_f 1^{\mathcal{A}} = a_{tr}$.

For the sake of simplicity, we write δ_2 for $\delta_2(X)$. Let us now apply *IROUGH* to $\mathbf{A}$ and $W(\delta_2) = X^{\leq 2}$. The computation is given in the table below. The current value of the set *NEW* is given in the corresponding row, while the values of S and $F(R^{\delta_2})$ include also all items appearing in the rows above the current row. For the sake of readability, we show even steps that add nothing to the sets S or *NEW*. In these steps, the algorithm just deletes the element of *NEW* considered. We always pick the first element from the list when an element from NEW has to be selected.

(a,x)	$Inv(a,x)$	S	NEW	$F(R^{\delta_2})$
		(a_f,ε)	(a_f,ε)	$\emptyset$
$(a_f,0)$	a_1	$(a_1,0)$	$(a_f,\varepsilon),(a_1,0)$	
$(a_f,1)$	$\emptyset$		$(a_1,0)$	
$(a_1,0)$	a_0	$(a_0,00)$	$(a_1,0),(a_0,00)$	00
$(a_1,1)$	a_1	$(a_1,10)$	$(a_0,00),(a_1,10)$	
$(a_0,0)$	$\emptyset$		$(a_0,00),(a_1,10)$	
$(a_0,1)$	$\emptyset$		$(a_1,10)$	
$(a_1,0)$	a_0	$(a_0,10)$	$(a_1,10),(a_0,10)$	10
$(a_1,1)$	a_1		$(a_0,10)$	
$(a_0,0)$	$\emptyset$		$(a_0,10)$	
$(a_0,1)$	$\emptyset$		$\emptyset$	

The set of final states obtained is $F(R^{\delta_2}) = \{00,10\}$, and it corresponds to the classes $[00]_{\delta_2},[10]_{\delta_2}$. Hence we obtain the approximation $R^{\delta_2} = X^*00 + X^*10$.

For the sake of comparison, we compute the same approximation R^{δ_2} using our first algorithm $ROUGH$.

x	b	S	NEW	$F(R^{\delta_2})$
		(a_0, ε)	(a_0, ε)	$\emptyset$
0	a_1	$(a_1, 0)$	$(a_0, \varepsilon), (a_1, 0)$	
1	a_{tr}	$(a_{tr}, 1)$	$(a_1, 0), (a_{tr}, 1)$	
0	a_f	$(a_f, 00)$	$(a_1, 0), (a_{tr}, 1), (a_f, 00)$	00
1	a_1	$(a_1, 01)$	$(a_{tr}, 1), (a_f, 00), (a_1, 01)$	
0	a_{tr}	$(a_{tr}, 10)$	$(a_{tr}, 1), (a_f, 00), (a_1, 01)$	
1	a_{tr}	$(a_{tr}, 11)$	$(a_f, 00), (a_1, 01), (a_{tr}, 11)$	
0	a_{tr}	$(a_{tr}, 00)$	$(a_f, 00), (a_1, 01), (a_{tr}, 11), (a_{tr}, 00)$	
1	a_{tr}	$(a_{tr}, 01)$	$(a_1, 01), (a_{tr}, 11), (a_{tr}, 00), (a_{tr}, 01)$	
0	a_f	$(a_f, 10)$	$(a_1, 01), (a_{tr}, 11), (a_{tr}, 00), (a_{tr}, 01), (a_f, 10)$	10
1	a_1	$(a_1, 11)$	$(a_{tr}, 11), (a_{tr}, 00), (a_{tr}, 01), (a_f, 10), (a_1, 11)$	
0	a_{tr}		$(a_{tr}, 11), (a_{tr}, 00), (a_{tr}, 01), (a_f, 10), (a_1, 11)$	
1	a_{tr}		$(a_{tr}, 00), (a_{tr}, 01), (a_f, 10), (a_1, 11)$	
0	a_{tr}		$(a_{tr}, 00), (a_{tr}, 01), (a_f, 10), (a_1, 11)$	
1	a_{tr}		$(a_{tr}, 01), (a_f, 10), (a_1, 11)$	
0	a_{tr}		$(a_{tr}, 01), (a_f, 10), (a_1, 11)$	
1	a_{tr}		$(a_f, 10), (a_1, 11)$	
0	a_{tr}		$(a_f, 10), (a_1, 11)$	
1	a_{tr}		$(a_1, 11)$	
0	a_f		$(a_1, 11)$	
1	a_1		$\emptyset$	

As the above table shows, many more steps are now needed because, in a sense, all paths from the initial state a_0 to the final state a_f are traversed. □

For any $k \geq 0$ and any X, the *reverse k-definite* languages ([14],[12],[38]) are defined by the relation $\rho_k(X)$ on X^+:

$$u\,\rho_k(X)\,v \;\Leftrightarrow\; \mathrm{pref}_k(u) = \mathrm{pref}_k(v) \quad (u, v \in X^+).$$

A language $L \subseteq X^+$ is *reverse k-definite* if it is saturated by $\rho_k(X)$. These languages are symmetric to k-definite in the sense that a language is reverse k-definite, if its reverse is k-definite. As a second example of application of the algorithm, we consider the generalized definite languages, that are a combination of k-definite and reverse h-definite. For any $h, k \geq 0$ and any X, we define the relation $\gamma_{h,k}(X)$ on X^+ by

$$u\,\gamma_{h,k}(X)\,v \;\Leftrightarrow\; \mathrm{pref}_h(u) = \mathrm{pref}_h(v) \text{ and } \mathrm{suff}_k(u) = \mathrm{suff}_k(v) \quad (u, v \in X^+).$$

A language $L \subseteq X^+$ is *h, k-definite* if it is saturated by $\gamma_{h,k}(X)$. For any pair $h, k \geq 0$, the (ε-free) h, k-definite languages form the principal $+$-variety corresponding to the principal $+$-filter $\{[\gamma_{h,k}(X))\}_X$. When *IROUGH* is used for computing least upper (h, k)-definite approximations, an appropriate set $W(\gamma_{h,k}(X))$ for representing the $\gamma_{h,k}(X)$-classes is $X^{\leq h} \times X^{\leq k}$ with $\widehat{w} = (\mathrm{pref}_h(w), \mathrm{suff}_k(w))$ for every $w \in X^*$.

3.3.6 Example If we apply *IROUGH* for finding the least upper 2, 2-definite approximation of the language $R = 01^*0$ over $X = \{0, 1\}$, we get $\{(00, 00), (01, 10)\}$ as the set of final states $F(R^{\gamma_{2,2}(X)})$, and hence the desired approximation is the language $R^{\gamma_{2,2}(X)} = 00 + 000 + 010 + 00X^*00 + 01X^*10$. Note that the words 00, 000 and 010 also belong to the classes represented by $(00, 00)$ or $(01, 10)$. □

As one more example, let us consider the locally testable languages. For any $k \geq 1$ and any alphabet X, we define the relation $\lambda_k(X)$ on X^+ by stipulating that for any $u, v \in X^+$, $u\,\lambda_k(X)\,v$ iff $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$, $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$ and $\mathrm{sw}_k(u) = \mathrm{sw}_k(v)$. It is easy to see that $kLoc^c := \{[\lambda_k(X))\}$ is a principal $+$-filter. A language $L \subseteq X^+$ is *k-testable* [29] if it is saturated by $\lambda_k(X)$, that is to say, if the membership of a word w in L is determined by its prefix of length $k-1$, its suffix of length $k-1$, and the set $\mathrm{sw}_k(w)$ of its subwords of length k. The family of (ε-free) k-testable languages $kLoc = \{kLoc(X)\}_X$ is the principal $+$-variety corresponding to the $+$-filter $kLoc^c$. Because $\lambda_k(X) \supset \lambda_{k+1}(X)$ for every $k \geq 0$ and any X, we have a properly ascending chain $Loc_0^c \subset 1Loc^c \subset 2Loc^c \subset \ldots$ of principal $+$-filters and a corresponding chain $0Loc \subset 1Loc \subset 2Loc \subset \ldots$ of principal $+$-varieties.

From the definition of $\lambda_k(X)$ it is clear that $W(\lambda_k(X)) := (X^{k-1}, \wp(X^k), X^{k-1})$ with $\widehat{w} = (\mathrm{pref}_{k-1}(w), \mathrm{sw}_k(w), \mathrm{suff}_{k-1}(w))$ is an appropriate representation of $X^*/\lambda_k(X)$

3.3.7 Example If we apply $IROUGH$ for computing the least upper 2-testable approximation of our example language $R = 01^*0$, we obtain as the set of final states

$$F(R^{\lambda_2(X)}) \;=\; \{(0,\{00\},0),\ (0,\{01,10\},0),\ (0,\{01,11,10\},0)\},$$

and hence we obtain the least upper 2-testable approximation

$$\begin{aligned} R^{\lambda_2(X)} &= \{u \mid u \in 0X^*0, \mathrm{sw}_2(u) \in \{\{00\},\{01,10\},\{01,11,10\}\}\} \\ &= 00^*0 + (0X^*0 \setminus X^*00X^*) \end{aligned}$$

□

3.4 Approximations in non-principal $+$-varieties

The most obvious non-principal $+$-variety is the greatest $+$-variety Rec of all the ε-free regular languages that corresponds to the greatest $+$-filter $Rec^c = \{\mathrm{FCon}(X^+)\}_X$. For any X and $R \in Rec(X)$, we get $R^{Rec} = R_{Rec} = R$. However, in general, the situation is more complicated.

Often a non-principal $+$-variety is naturally defined as the union of an ascending chain of principal $+$-varieties or, more generally, as the union of a directed family of principal $+$-varieties; recall that a non-empty family of classes $\mathcal{S}$ is *directed* if for all $A, B \in \mathcal{S}$, there is a $C \in \mathcal{S}$ such that $A, B \subseteq C$. On the other hand, it is easy to see that the union of a directed family of principal $+$-varieties is always a $+$-variety.

3.4.1 Lemma *Let $\mathcal{L}$ be the union $\bigcup_{i\in I} \mathcal{L}_i$ of a directed family of principal $+$-varieties*

$\mathcal{L}_i (i \in I)$. For any alphabet X and any language $R \subseteq X^+$, if $R^{\mathcal{L}}$ exists, then $R^{\mathcal{L}} = R^{\mathcal{L}_i}$ for some $i \in I$. Similarly, if $R_{\mathcal{L}}$ exists, then $R_{\mathcal{L}} = R_{\mathcal{L}_i}$ for some $i \in I$.

Proof. Let $\mathcal{L}_i^c = \{[\theta_i(X))\}_X$ $(i \in I)$. If $R^{\mathcal{L}}$ exists, then by Corollary 3.2.9 $R^{\mathcal{L}} = R^{\theta}$ for some $\theta \in \mathcal{L}^c(X)$. On the other hand, it is easy to see that $\mathcal{L}^c(X) = [\{\theta_i(X) \mid i \in I\})$, and since $\{\theta_i(X) \mid i \in I\}$ is (downwards) directed, we must have $\theta \supseteq \theta_i(X)$ for some $i \in I$, and then $R \subseteq R^{\theta_i(X)} \subseteq R^{\theta}$ implies that $R^{\mathcal{L}} = R^{\theta_i(X)} = R^{\mathcal{L}_i}$. The second fact is obtained dually. □

Let us consider an example.

3.4.2 Example A language is *definite* [17, 36] if it is k-definite for some $k \geq 0$. Hence the $+$-variety Def of definite (ε-free) languages is the union of the ascending chain $0Def \subset 1Def \subset 2Def \subset \ldots$ of principal $+$-varieties, and the corresponding $+$-filter Def^c is the union of the chain $0Def^c \subset 1Def^c \subset 2Def^c \subset \ldots$ of principal $+$-filters.

Let us consider the non-definite language $R = 0^*10^*$ over $X = \{0, 1\}$. By Example 3.3.3, for each $k \geq 0$,

$$R^{kDef} = \{0^i10^j \mid i, j \geq 0, i + j < k - 1\} \cup \bigcup \{X^*w \mid w \in X^k, R \cap X^*w \neq \emptyset\},$$

and thus $0^k \in R^{kDef} \setminus R^{(k+1)Def}$. Hence we obtain the properly descending chain

$$X^+ = R^{0Def} \supset R^{1Def} \supset R^{2Def} \supset \ldots (\supset R)$$

of upper approximations of R. Thus none of the approximations R^{kDef} is the least upper Def-approximation of R, and by Lemma 3.4.1 this means that R^{Def} does not exist. On the other hand, for every $k \geq 0$, we have $R_{kDef} = \{0^i10^j \mid i, j \geq 0, i + j < k - 1\}$, and hence $10^{k-1} \in R_{(k+1)Def} \setminus R_{kDef}$. This means that $R_{0Def} \subset R_{1Def} \subset R_{2Def} \subset \ldots (\subset R)$, and thus R has no greatest lower Def-approximation either. □

The following proposition shows that, in fact, the closest definite approximations R^{Def} and R_{Def} do not exist for any non-definite language R.

3.4.3 Proposition *Let $\mathcal{L}$ be a $+$-variety and X be any alphabet. If $\{w\} \in \mathcal{L}(X)$ for every $w \in X^+$, then the least upper $\mathcal{L}$-approximation $R^{\mathcal{L}}$ of a language $R \subseteq X^+$ exists iff $R \in \mathcal{L}(X)$. Similarly, $R_{\mathcal{L}}$ exists iff $R \in \mathcal{L}(X)$.*

Proof. If $R \in \mathcal{L}(X)$, then naturally $R^{\mathcal{L}} = R$. Assume then that $R \notin \mathcal{L}(X)$ and let $L \in \mathcal{L}(X)$ be any upper $\mathcal{L}$-approximation of R. For any $w \in L - R\ (\neq \emptyset)$, also $L - \{w\}$ is an upper $\mathcal{L}$-approximation of R, and hence $R^{\mathcal{L}}$ cannot exist. The second assertion is proved similarly. □

It is easy to see that if a $+$-variety $\mathcal{L}$ is the union $\bigcup_{i\in I} \mathcal{L}_i$ of a directed family of principal $+$-varieties $\mathcal{L}_i$, where $\mathcal{L}_i^c = \{[\theta_i(X))\}_X$ $(i \in I)$, then for any X and $w \in X^+$, $\{w\} \in \mathcal{L}(X)$ iff $[w]_{\theta_i(X)} = \{w\}$ for some $i \in I$. Hence, we may restate the above proposition for a frequently occurring case as follows.

3.4.4 Corollary *Let $\mathcal{L} = \bigcup_{i\in I} \mathcal{L}_i$ be a $+$-variety given as the union of a directed family of principal $+$-varieties, where $\mathcal{L}_i^c = \{\left[\theta_i(X)\right)\}_X$ $(i \in I)$, and assume that for some alphabet X, there exists for every word $w \in X^+$ an $i \in I$ such that $[w]_{\theta_i(X)} = \{w\}$. Then a language $R \subseteq X^+$ has a least upper $\mathcal{L}$-approximation iff $R \in \mathcal{L}(X)$. Similarly, $R_{\mathcal{L}}$ exists iff $R \in \mathcal{L}(X)$.* □

As we noted in Examples 3.3.3 and 3.3.4, $[w]_{\delta_k(X)} = \{w\}$ and $[w]_{J_k(X)} = \{w\}$ for any $k \geq 0$ and $w \in X^{<k}$, and hence Corollary 3.4.4 applies to the $+$-varieties of Def and $\mathcal{J}$. It also applies, for example, to the non-principal $+$-varieties of *finite and co-finite* languages, *reverse definite* languages, *generalized definite* languages, and *locally testable* languages ([12, 6, 14, 29, 38]). A language has a least upper or greatest lower approximation in any of those $+$-varieties only in case it itself belongs to the variety. Membership to each of these families can be decided by inspecting the syntactic semigroup of the language (cf. [12, 38]). Let us conclude this section with a couple of examples. The first one shows that in a non-principal $+$-variety to which Proposition 3.4.3 does not apply, the closest approximations may exist even for languages that do not belong to the variety.

3.4.5 Example For any word w, let $c(w)$ denote the set of all words that can be obtained from w by permuting its letters. A language R is *commutative* if $c(w) \subseteq R$ for every $w \in R$. Let $Com(X)$ denote the set of ε-free regular commutative languages over X. The family $Com := \{Com(X)\}_X$ is a $+$-variety, cf. [12, 38]. Clearly, the assumption of Proposition 3.4.3 cannot hold for Com and any X with at least two letters. The non-commutative language $R = \{00, 10\}$ over the alphabet $\{0, 1\}$ has both of the closest Com-approximations. Indeed, $R^{Com} = \{00, 01, 10\}$ and $R_{Com} = \{00\}$. □

The second example shows that in a non-principal $+$-variety $\mathcal{L}$ a language R may have one of the closest approximations $R^{\mathcal{L}}$ and $R_{\mathcal{L}}$ without having the other.

3.4.6 Example Let $\mathcal{L} = Com \cap Def$ be the $+$-variety of commutative definite ε-free languages. For each $k \geq 0$, let $\mathcal{L}_k = Com \cap kDef$. Then $\mathcal{L} = \bigcup_{k \geq 0} \mathcal{L}_k$ and $\mathcal{L}_0 \subset \mathcal{L}_1 \subset \mathcal{L}_2 \subset \ldots$. It is easy to see that for every $k \geq 0$, $\mathcal{L}_k$ is the principal $+$-variety defined by the principal $+$-filter $\mathcal{L}_k^c = \{[\theta_k(X))\}_X$ where $\theta_0(X) = \nabla_{X^+}$ and for any $k \geq 1$,

$$X^+/\theta_k(X) = \{c(w) \mid 1 \leq \lg(w) < k\} \cup \{X^{\geq k}\};$$

any two words $u, v \in X^{\geq k}$ are $\theta_k(X)$-related because $(u, vu), (vu, uv), (uv, v) \in \theta_k(X)$. Hence, if $R^{\mathcal{L}}$ (or $R_{\mathcal{L}}$) exists, then $R^{\mathcal{L}} = R^{\theta_k(X)}$ (or $R_{\mathcal{L}} = R_{\theta_k(X)}$) for some $k \geq 0$.

Let us consider the regular language $R = (01)^+ = \{01, 0101, \ldots\}$ over $X = \{0, 1\}$. Since $R_{\theta_k} = \emptyset$ for every $k \geq 0$, we have $R_{\mathcal{L}} = \emptyset$. On the other hand, $R^{\theta_0(X)} = X^+$ and

$$R^{\theta_k(X)} = \bigcup \{c((01)^i) \mid 1 \leq i < k/2\} \cup X^{\geq k}$$

for each $k \geq 1$, and hence $R^{\theta_0(X)} \supset R^{\theta_1(X)} \supset R^{\theta_2(X)} \supset \ldots$ and $R^{\mathcal{L}}$ cannot exist. For the complement R' the converse holds; $(R')^{\mathcal{L}} = X^+$ exists but $(R')_{\mathcal{L}}$ does not. □

3.5 Approximations in pseudo-principal $+$-varieties

Let us now consider approximations in certain $+$-varieties that are natural generalizations of principal $+$-varieties.

We call a system $\beta = \{\beta(X)\}_X$ a *family of congruences* if $\beta(X) \in \mathrm{Con}(X^+)$ for each alphabet X. Such a family is said to be *consistent* if $\beta(X) \subseteq \varphi \circ \beta(Y) \circ \varphi^{-1}$ for all alphabets X, Y and every homomorphism $\varphi : X^+ \to Y^+$. For any family of congruences $\beta = \{\beta(X)\}_X$, let $\Gamma_\beta = \{\Gamma_\beta(X)\}_X$, where $\Gamma_\beta(X) := \{\theta \in \mathrm{FCon}(X^+) \mid \beta(X) \subseteq \theta\}$ for each alphabet X. Let us note a few basic properties of these notions.

3.5.1 Lemma *Let $\beta = \{\beta(X)\}_X$ be a consistent family of congruences.*

(a) *For every X, $\beta(X)$ is a fully invariant congruence of X^+.*

(b) *If $|X| = |Y|$, then $X^+/\beta(X) \cong Y^+/\beta(Y)$.*

(c) Γ_β *is a $+$-filter.*

(d) *If $\beta(X) \in \mathrm{FCon}(X^+)$ for every X, then Γ_β equals the principal $+$-filter $\{[\beta(X))\}_X$.*

Proof. For (a), it suffices to apply the consistency condition to the endomorphisms $\varphi : X^+ \to X^+$. To prove (b), let $\psi_0 : X \to Y$ be a bijection and let $\psi : X^+ \to Y^+$ be its extension to an isomorphism. Then one can easily verify that $[w]_{\beta(X)} \mapsto [w\psi]_{\beta(Y)}$ yields a well-defined isomorphism $X^+/\beta(X) \to Y^+/\beta(Y)$.

Clearly, $\Gamma_\beta(X)$ is a filter of $\mathrm{FCon}(X^+)$ for every X. Consider any $\theta \in \Gamma_\beta(Y)$ and any homomorphism $\varphi : X^+ \to Y^+$. Then $\theta \in \mathrm{FCon}(Y^+)$ implies $\varphi \circ \theta \circ \varphi^{-1} \in \mathrm{FCon}(X^+)$, and by the consistency condition, $\beta(X) \subseteq \varphi \circ \beta(Y) \circ \varphi^{-1} \subseteq \varphi \circ \theta \circ \varphi^{-1}$. Hence, also $\varphi \circ \theta \circ \varphi^{-1} \in \Gamma_\beta(X)$ holds, and (c) follows. Now (d) is quite obvious; if $\beta(X) \in \mathrm{FCon}(X)$, then $\Gamma_\beta(X) = [\beta(X))$. □

3.5.2 Definition We call a $+$-filter $\Gamma = \{\Gamma(X)\}_X$ *pseudo-principal* if $\Gamma = \Gamma_\beta$ for some consistent family of congruences β, and then Γ^l is a *pseudo-principal $+$-variety.*

We shall now show that pseudo-principal $+$-varieties correspond, in the sense of Eilenberg's [12] Variety Theorem, to so-called equational varieties of finite semigroups.

First we recall some notions and facts concerning equational classes and varieties of finite semigroups. For systematic treatments of these matters, cf. [1], [7], [12], [38] or [39], for example.

A class of semigroups $\mathbf{V}$ is a *variety* if it contains all subsemigroups, all homomorphic images and all direct products of members of $\mathbf{V}$. If $\mathbf{K}$ is any class of semigroups, the *variety generated* by $\mathbf{K}$, i.e., the least variety $\mathbf{V}$ such that $\mathbf{K} \subseteq \mathbf{V}$, is denoted by $\mathrm{V}(\mathbf{K})$. Moreover, let $\mathrm{F}(\mathbf{K})$ be the class of all finite members of $\mathbf{K}$.

An *identity* over an alphabet Z, not necessarily finite, is an expression $u \approx v$ where $u, v \in Z^+$. A semigroup S *satisfies* $u \approx v$, $S \models u \approx v$ in symbols, if $u\varphi = v\varphi$ for every homomorphism $\varphi : Z^+ \to S$. If E is a set of identities, $S \models E$ means that $S \models u \approx v$ for every $u \approx v \in E$. More generally, a class $\mathbf{K}$ of semigroups *satisfies* $u \approx v$ (resp., E), and we write $\mathbf{K} \models u \approx v$ ($\mathbf{K} \models E$), if $S \models u \approx v$ ($S \models E$) for every $S \in \mathbf{K}$. It is convenient to identify an identity $u \approx v$ with the ordered pair (u, v) of words. Then the set of identities over Z satisfied by a class $\mathbf{K}$ of semigroups equals the fully invariant congruence $\theta_{\mathbf{K}}(Z)$ of Z^+ that is defined as the intersection of the kernels $\ker\varphi$, where $\varphi : Z^+ \to S$ is a homomorphism for some $S \in \mathbf{K}$ (cf. [7], especially Sections II.11 and II.14, or [1], for example). On the other hand, for any set of identities E, let $\mathrm{Mod}(E)$ be the class of all the semigroups that satisfy all the identities of E. A class $\mathbf{K}$ of semigroups is called *equational* if $\mathbf{K} = \mathrm{Mod}(E)$ for some set E of identities. A fundamental theorem of G. Birkhoff (for general algebras) states that a class is equational iff it is a variety.

The following fact has a straightforward proof.

3.5.3 Lemma *For every class $\mathbf{K}$ of semigroups, $\theta_{\mathbf{K}} := \{\theta_{\mathbf{K}}(X)\}_X$ is a consistent family of congruences.* □

For the next lemma, cf. Lemma II.14.7 in [7], for example.

3.5.4 Lemma *If $\rho \in \mathrm{Con}(X^+)$ is fully invariant, then for all $u, v \in X^+$, $X^+/\rho \models u \approx v$ iff $u \,\rho\, v$.* □

In the following lemma, the congruence ρ is viewed also as a set of identities.

3.5.5 Lemma *Let $\rho, \theta \in \mathrm{Con}(X^+)$ for some X. If ρ is fully invariant, then $X^+/\theta \models \rho$ iff $\rho \subseteq \theta$.*

Proof. Recall that, for any congruence θ of a semigroup S, the natural mapping $\nu_\theta : S \to S/\theta,\ s \mapsto [s]_\theta$, is a homomorphism. Hence, if $X^+/\theta \models \rho$, then $[u]_\theta = u\nu_\theta = v\nu_\theta = [v]_\theta$ for any $(u, v) \in \rho$, i.e., $\rho \subseteq \theta$.

If $\rho \subseteq \theta$, then $X^+/\rho \to X^+/\theta,\ [w]_\rho \mapsto [w]_\theta$, is a well-defined epimorphism, and since $X^+/\rho \models \rho$ by Lemma 3.5.4, also $X^+/\theta \models \rho$ holds. □

Various forms of the following lemma are known in general algebra but for the sake of completeness, we prove it as stated.

3.5.6 Lemma *Let $\mathbf{V}$ be a variety of semigroups, X be an alphabet and $\theta \in \mathrm{Con}(X^+)$. Then $X^+/\theta \in \mathbf{V}$ iff $\theta_{\mathbf{V}}(X) \subseteq \theta$.*

Proof. If $\theta_{\mathbf{V}}(X) \subseteq \theta$, then $X^+/\theta_{\mathbf{V}}(X) \to X^+/\theta,\ [w]_{\theta_{\mathbf{V}}(X)} \mapsto [w]_\theta$, is an epimorphism. Moreover, $X^+/\theta_{\mathbf{V}}(X) \in \mathbf{V}$ since $\mathbf{V}$ is a variety (cf. Corollary II.11.10 of [7]). This means that $X^+/\theta \in \mathbf{V}$, too. On the other hand, $X^+/\theta \in \mathbf{V}$ implies that $X^+/\theta \models \theta_{\mathbf{V}}(X)$, and hence $\theta_{\mathbf{V}}(X) \subseteq \theta$ by Lemma 3.5.5. □

A *variety of finite semigroups*, a *VFS* for short, is a class $\mathbf{S}$ of finite semigroups that contains all subsemigroups, all homomorphic images and all finite direct products of its members. For any variety $\mathbf{V}$ of semigroups, the subclass $\mathrm{F}(\mathbf{V})$ is a variety of finite semigroups, and a VFS $\mathbf{S}$ is called *equational* if $\mathbf{S} = \mathrm{F}(\mathbf{V})$ for some variety $\mathbf{V}$ (cf. [1], p. 60). Hence, a class $\mathbf{S}$ of finite semigroups is an equational VFS iff there exists a set E of identities (over some alphabet) such that $\mathbf{S} = \mathrm{F}(\mathrm{Mod}(E))$.

Let $\mathbf{S}$ be any VFS. For each (finite) alphabet X, let $\mathbf{S}^l(X)$ consist of all the languages $L \subseteq X^+$ such that its *syntactic semigroup* $S(L) := X^+/\sigma_L$ is in $\mathbf{S}$. Then $\mathbf{S}^l := \{\mathbf{S}^l(X)\}_X$ is the $+$-variety that by the Variety Theorem corresponds to $\mathbf{S}$. Conversely, for any $+$-variety $\mathcal{L} = \{\mathcal{L}(X)\}_X$, the corresponding VFS $\mathcal{L}^s$ is the VFS generated by the syntactic semigroups $S(L)$ with $L \in \mathcal{L}(X)$ for some X.

The $+$-filter $\mathbf{S}^c = \{\mathbf{S}^c(X)\}_X$ that by Thérien's theorem corresponds to a given VFS $\mathbf{S}$ is defined by $\mathbf{S}^c(X) := \{\theta \in \mathrm{FCon}(X^+) \mid X^+/\theta \in \mathbf{S}\}$. Conversely, the VFS

Γ^s that corresponds to a given $+$-filter $\Gamma = \{\Gamma(X)\}_X$, is the VFS generated by the semigroups X^+/θ, where $\theta \in \Gamma(X)$ for some X. Lemma 3.5.6 yields the following description of the $+$-filter corresponding to an equational VFS.

3.5.7 Lemma *If* $\mathbf{V}$ *is any variety of semigroups, then* $\mathrm{F}(\mathbf{V})^c = \Gamma_{\theta_{\mathbf{V}}}$*, that is to say,* $\mathrm{F}(\mathbf{V})^c(X) = \{\theta \in \mathrm{FCon}(X^+) \mid \theta_{\mathbf{V}}(X) \subseteq \theta\}$ *for every alphabet* X*.* □

For any family of congruences $\beta = \{\beta(X)\}_X$, let $\mathbf{K}(\beta)$ be the class of all quotient semigroups $X^+/\beta(X)$, where X ranges over all (finite) alphabets, and let $\mathbf{V}(\beta)$ be the variety generated by $\mathbf{K}(\beta)$.

3.5.8 Lemma *If* β *is a consistent family of congruences, then* $\theta_{\mathbf{V}(\beta)} = \beta$*.*

Proof. Consider any alphabet X. Since $\theta_{\mathbf{V}(\beta)}(X) = \theta_{\mathbf{K}(\beta)}(X)$, it suffices to show that $\theta_{\mathbf{K}(\beta)}(X) = \beta(X)$. The inclusion $\theta_{\mathbf{K}(\beta)}(X) \subseteq \beta(X)$ follows from Lemma 3.5.4. Indeed, if $(u, v) \in \theta_{\mathbf{K}(\beta)}(X)$, then $X^+/\beta(X) \models u \approx v$, and hence $(u, v) \in \beta(X)$.

To prove the converse inclusion, let $(u, v) \in \beta(X)$. Consider any alphabet Y and any homomorphism $\psi : X^+ \to Y^+/\beta(Y)$. There is a homomorphism $\varphi : X^+ \to Y^+$ such that $\psi = \varphi \circ \nu_{\beta(Y)}$. Then the consistency condition $\beta(X) \subseteq \varphi \circ \beta(Y) \circ \varphi^{-1}$ implies that $u\varphi\, \beta(Y)\, v\varphi$, and therefore $u\psi = (u\varphi)\nu_{\beta(Y)} = (v\varphi)\nu_{\beta(Y)} = v\psi$. This shows that $Y^+/\beta(Y) \models u \approx v$, and hence $\beta(X) \subseteq \theta_{\mathbf{K}(\beta)}(X)$ holds, too. □

Now we can establish the following correspondence.

3.5.9 Proposition *A variety of finite semigroups* $\mathbf{S}$ *is equational iff the corresponding* $+$*-filter* $\mathbf{S}^c$ *is pseudo-principal.*

Proof. If $\mathbf{S} = \mathrm{F}(\mathbf{V})$ for some variety $\mathbf{V}$ of semigroups, $\theta_{\mathbf{V}}$ is a consistent family of congruences by Lemma 3.5.3, and $\mathbf{S}^c = \Gamma_{\theta_{\mathbf{V}}}$ by Lemma 3.5.7. Assume then that $\mathbf{S}^c = \Gamma_\beta$ for a consistent family β of congruences. Then $\mathbf{S}^c = \Gamma_\beta = \Gamma_{\theta_{\mathbf{V}(\beta)}} = \mathrm{F}(\mathbf{V}(\beta))^c$ by Lemmas 3.5.7 and 3.5.8, and hence $\mathbf{S}$ is the equational VFS $\mathrm{F}(\mathbf{V}(\beta))$. □

3.5.10 Lemma *If* $(R^\theta)^\rho = R^\theta$ *for some* $R \subseteq X^+$ *and* $\theta, \rho \in \mathrm{Eq}(X^+)$*, then* $R^{\theta\vee\rho} = R^\theta$*.*

Proof. The assertion follows from Lemma 2.0.1 (b); since R^θ is in both $\mathrm{Sat}(\theta)$ and $\mathrm{Sat}(\rho)$, we have $R^\theta \in \mathrm{Sat}(\theta \vee \rho)$, and hence $R^\theta \subseteq R^{\theta\vee\rho} \subseteq (R^\theta)^{\theta\vee\rho} = R^\theta$. □

3.5.11 Proposition *Let β be a consistent family of congruences and let $\mathcal{L} = \Gamma^l_\beta$ be the corresponding $+$-variety. For any alphabet X and any language $R \subseteq X^+$, the following three conditions are pairwise equivalent.*

(1) *The upper approximation $R^{\beta(X)}$ is a regular language.*

(2) *$R^{\beta(X)} = R^\theta$ for some $\theta \in \Gamma_\beta(X)$.*

(3) *$R^{\beta(X)} = R^{\mathcal{L}}$.*

Proof. If $R^{\beta(X)}$ is regular, then $R^{\beta(X)} = (R^{\beta(X)})^\rho$ for some $\rho \in \mathrm{FCon}(X^+)$. Now $R^{\beta(X)} = R^{\beta(X)\vee\rho}$ by Lemma 3.5.10, and naturally $\beta(X) \vee \rho \in \Gamma_\beta(X)$. Hence (1) implies (2), and the converse is obvious.

To show that (2) implies (3), let $R^{\beta(X)} = R^\theta$ for some $\theta \in \Gamma_\beta(X)$. Then $R^{\beta(X)}$ is an upper $\mathcal{L}$-approximation of R, and from Proposition 3.2.8 and Lemma 3.1.3 it easily follows that it is the least upper $\mathcal{L}$-approximation.

Finally, if $R^{\beta(X)} = R^{\mathcal{L}}$, then it follows from Corollary 3.2.9 that $R^{\beta(X)} = R^\theta$ for some $\theta \in \Gamma_\beta(X)$. Hence, (3) implies (2). □

We say that $\theta \in \mathrm{Con}(X^+)$ has *regular classes* if $[u]_\theta \in Rec(X)$ for every $u \in X^+$.

3.5.12 Lemma *Let β be a consistent family of congruences and let $u \in X^+$ for some alphabet X. Then $[u]_{\beta(X)}$ is a regular language iff there exists a congruence $\rho \in \Gamma_\beta(X)$ such that $[u]_{\beta(X)} = [u]_\rho$.*

Proof. If $[u]_{\beta(X)}$ is a regular language, it is saturated by some $\theta \in \mathrm{FCon}(X^+)$. Now, $\rho := \theta \vee \beta(X) \in \Gamma_\beta(X)$ and $[u]_{\beta(X)} = [u]_\rho$ by Lemma 3.5.10. This implies one direction of the lemma, and the converse is perfectly obvious. □

3.5.13 Proposition *Let β be a consistent family of congruences and let $\mathcal{L}$ be the corresponding $+$-variety. Furthermore, assume that $\beta(X)$ has regular classes for some X. For any $R \subseteq X^+$, if $R^{\mathcal{L}}$ exists, then $R^{\mathcal{L}} = R^{\beta(X)}$.*

Proof. If $R^{\mathcal{L}}$ exists, then $R^{\mathcal{L}} = R^{\theta}$ for some $\theta \in \Gamma_{\beta}(X)$. If $R^{\beta(X)} \subset R^{\theta}$, then there is a word $u \in R^{\theta}$ such that $(u, v) \in \beta(X)$ for no $v \in R$. If $\rho \in \Gamma_{\beta}(X)$ is the congruence prescribed by Lemma 3.5.12 for this u, then $R^{\theta\cap\rho} \subset R^{\theta}$ since $u \in R^{\theta} \setminus R^{\rho\cap\theta}$. As $\theta \cap \rho \in \Gamma_{\beta}(X)$, this would mean that R^{θ} is not the least upper $\mathcal{L}$-approximation of R. Hence $R^{\beta(X)} = R^{\theta} = R^{\mathcal{L}}$ must hold. □

Applied to a pseudo-principal $+$-variety given by the corresponding equational VFS, the above findings can be summarized as follows.

3.5.14 Corollary *Let $\mathcal{L} = \{\mathcal{L}(X)\}_X$ be the pseudo-principal $+$-variety that corresponds to a given equational VFS* $\mathrm{F}(\mathbf{V})$*, where* $\mathbf{V}$ *is a variety of semigroups, and let $R \subseteq X^+$ for some alphabet X.*

(a) *If $R^{\theta_{\mathbf{V}}(X)}$ is a regular language, then it is the least upper $\mathcal{L}$-approximation of R.*

(b) *If $\theta_{\mathbf{V}}(X)$ has regular classes, then the least upper $\mathcal{L}$-approximation of R exists iff $R^{\theta_{\mathbf{V}}(X)}$ is a regular language, and then $R^{\mathcal{L}} = R^{\theta_{\mathbf{V}}(X)}$.* □

Let us consider, by the way of a concrete example, the $+$-variety *Com* of commutative languages. For any alphabet X, let $\varkappa(X)$ be the equivalence on X^+ such that $X^+/\varkappa(X) = \{c(w) \mid w \in X^+\}$. Then $\varkappa(X)$ is a congruence on X^+, and a language $R \subseteq X^+$ is commutative iff it is saturated by $\varkappa(X)$. It is obvious that the family of congruences $\varkappa = \{\varkappa(X)\}_X$ is consistent and that *Com* is the pseudo-principal $+$-variety defined by $\Gamma_{\varkappa}$, i.e., that $Com^c(X) = \{\theta \in \mathrm{FCon}(X^+) \mid \varkappa(X) \subseteq \theta\}$ for every alphabet X. It is well known (cf. [12, 38]) that the corresponding equational VFS is $\mathrm{F}(\mathbf{Com})$, where $\mathbf{Com}$ is the variety of commutative semigroups. Hence, $\theta_{\mathbf{Com}} = \varkappa$. Since $[w]_{\varkappa(X)} = c(w)$ for any X and $w \in X^+$, the family of congruences $\varkappa$ has regular classes and $R^{\varkappa(X)} = c(R) := \bigcup\{c(w) \mid w \in R\}$ for any $R \subseteq X^+$. By Corollary 3.5.14, the least upper *Com*-approximation exists iff the commutative closure $c(R)$ of R is regular. For example, a regular language like $(01)^+$ does not have a least upper *Com*-approximation. However, since it is decidable whether the commutative closure of a regular language R is regular (cf. [13], for example), it is decidable whether R^{Com}

exists. When $c(R)$ is non-regular, one has to be satisfied with other commutative approximations based on some congruence θ in $\Gamma_{\varkappa}$.

3.6 Concluding remarks

We have introduced certain approximations of languages and presented a number of their basic properties. Of course, much remains to be done. For example, for certain $+$-varieties it could be possible to develop methods to efficiently find the approximations by making use of the special properties of the variety.

Pseudo-principal $+$-varieties pose several natural and challenging problems. When and how can we decide whether the closest approximations exist? How can they be formed when they exist? And if they don't exist, how can we find other approximations in the variety that are sufficiently close to the given language?

Although the most general notions and results apply directly also to other types of varieties, such as varieties of tree languages, new problems will probably be encountered in the more advanced parts of the theory.

Chapter 4

Accuracy of rough approximations in a $+$-variety

Throughout this chapter we consider the accuracy of upper θ-approximations R^θ of a given language R over an alphabet X with respect to a given congruence. We consider only upper rough approximations because, as we will observe in this chapter, languages often do not contain whole classes of the given congruence thus giving empty lower rough approximations.

Several ways to measure the quality of rough approximations have been proposed (see for example [35, 32, 33]), but they are not useful in our case because they deal with finite sets of objects. We adopt the approach suggested by Berstel in [4]. He introduced an expression for the relative density of two given languages, one of which is a subset of the other. Here we consider the density of a language in its upper approximation; this number, when defined, is regarded as a measure of the accuracy of the approximation. In particular, we look at the accuracy of approximations in the families of k-definite, reverse k-definite, generalized i, j-definite, k-testable and commutative languages. Some of the results are new, but most come from our paper [28], unless stated otherwise. When the alphabet is clear from the context, we write just θ instead of $\theta(X)$.

4.1 Densities of the languages in the varieties under study

The following standard definitions can be found in [18] or [5]. Given two functions $f(n)$ and $g(n)$, we say that $f(n)$ is $O(g(n))$, or that $f(n) = O(g(n))$, if for some positive constants c and n_0, $0 \leqslant f(n) \leqslant cg(n)$ for all $n \geqslant n_0$; $f(n)$ is said to be $\Omega(g(n))$, and written $f(n) = \Omega(g(n))$, if there exists a constant $c > 0$ and an infinite sequence $n_1 < n_2 < n_3 < ...$ satisfying $f(n_i) \geqslant cg(n_i)$ for all $i \geqslant 1$, and finally $f(n)$ is $\Theta(g(n))$, and written $f(n) = \Theta(g(n))$, if $f(n)$ is both $\Omega(g(n))$ and $O(g(n))$. The *density function* $d_R(m)$ of a language $R \subseteq X^+$ is defined by $d_R(m) = |R \cap X^m|$, $m \geq 1$. If $d_R(m) = \Theta(1)$ then we say that R has *constant density*, if $d_R(m) = \Theta(m^c)$ for some integer $c \geq 1$, then R has *polynomial density* and if $d_R(m) = \Theta(c^m)$ for some $c \geqslant 0$, then R has *exponential density.* Languages of constant density are called *slender languages.* Languages that have at most one word of each length are called *thin languages.*

In [47] the regular languages of polynomial density are characterized as unions of finitely many languages of the form $z_0y_1^*z_1...y_{k+1}^*z_{k+1}$ where $z_0, y_1, z_1, ..., y_{k+1}, z_{k+1} \in X^*$. Furthermore, it is shown that for every infinite regular language R, the density $d_R(m)$ is either polynomial or exponential of the form $2^{\Theta(cm)}$, where c is a constant.

We will now compute the densities of languages in the varieties under study.

It is easy to see that infinite $k-$definite, reverse $k-$definite, and generalized $i, j-$definite languages over an at least binary alphabet, have exponential density. Commutative and $k-$testable languages can have exponential, polynomial or constant density. The case of commutative languages is different from the others, as we will see along this chapter, partly because of the fact that its associated congruence $\varkappa$ has infinite index. Furthermore, approximations under this congruence are not always regular. In the next two propositions we describe some features of the densities of these families.

4.1.1 Proposition *Let X be an at least binary alphabet, and $R \subseteq X^+$ an infinite language in $Sat(\theta)$ for some $\theta \in \{\delta_k, \rho_k, \gamma_{i,j}, \lambda_k\}$, where $k, i+j \geqslant 1$.*

(1) *If $\theta \in \{\delta_k, \rho_k\}$ there is a number $1 \leqslant n \leqslant |X|^k$ such that $d_R(m) = n|X|^{m-k}$ for every $m \geqslant k$.*

(2) *If $\theta = \gamma_{i,j}$, there is a number $1 \leqslant n \leqslant |X|^{i+j}$ such that $d_R(m) = n|X|^{m-(i+j)}$ for every $m \geqslant i + j$.*

(3) *If $\theta = \lambda_k$, then R can have constant, exponential or polynomial density, and it is decidable which one it has. Furthermore, for every integer $n > 0$ there exists an alphabet X and a language $R \subseteq X^+$ in this family such that $d_R(m) = \Theta(m^n)$.*

Proof. If R is an infinite language in $Sat(\delta_k)$, there must be words $v_1, ..., v_r \in X^{<k}$, $r \geqslant 0$, and $u_1, ..., u_n \in X^k$ for some $n \geqslant 1$ such that $R = \{v_1, ..., v_r\} \cup X^*u_1 \cup ... \cup X^*u_n$. Thus $d_R(m) = n|X|^{m-k}$ for every $m \geqslant k$. When $\theta \in \{\rho_k, \gamma_{i,j}\}$ the proof is similar. Consider next the case of $k-$testable languages. The language $R = (01)^+$ is an example of slender $2-$testable. To see that there are exponential density $k-$testable languages, it suffices to note that X^+ is a $k-$testable language. For a (slightly) less trivial example, consider a $k-$testable language R over the alphabet $X = \{0, 1, 2, 3\}$ such that its not allowed subwords are in $\{0, 1\}^+$. Any language with this property contains $(2 + 3)^+$ as a subset, therefore it has exponential density.

We show now that there exist polynomial density $2-$testable languages. We start by considering the language $R = 1^*23^*45^*$ over the alphabet $X = \{1, 2, 3, 4, 5\}$. It is easy to see that R is a $2-$testable language over X. For each number $m \geqslant 2$, there is exactly one word in R of length m for each choice of the positions of 2 and 4. Since there are $\binom{m}{2}$ such choices, $d_R(m) = \binom{m}{2} = \frac{m(m-1)}{2}$ and this is $\Theta(m^2)$. Using this idea, we build now a $2-$testable language R such that $d_R(m) = \Theta(m^n)$ for any given $n \geqslant 1$. If $n = 1$ we take the language $R = x_0^*x_1x_2^*$ over the alphabet $X = \{x_0, x_1, x_2\}$ that satisfies $d_R(m) = m$. If $n > 1$ we take the alphabet $X = \{x_0, ..., x_{k+1}\}$ where k satisfies $2n = k + 1$, and the $2-$testable language $R = x_0^*x_1x_2^*x_3...x_kx_{k+1}^*$. In this case $d_m(R) = \binom{m}{n} = \Theta(m^n)$. Finally, in [47] and [43] regular languages of exponential density are characterized, and this characterization leads to a linear algorithm for deciding whether a regular language is of exponential density when the language is given by a deterministic finite automaton. □

We denote by $\psi(L)$ the Parikh set of a language $L \subseteq X^+$.

4.1.2 Proposition *Let X be an alphabet with l letters, and $L \subseteq X^+$ a language in $Sat(\varkappa)$.*

(1) $d_L(m) = \sum\{\binom{m}{s_1,\ldots,s_l} \mid (s_1,\ldots,s_l) \in \psi(L \cap X^m)\}$.

(2) *L can have constant, exponential or polynomial density, and when L is regular, it is decidable which one it has. Furthermore, for every integer $n > 0$ there exists a commutative language L, over a binary alphabet, such that $d_L(m) = \Theta(m^n)$.*

Proof. The expression of the density of a commutative language follows straightforwardly by counting the non-equivalent permutations of letters corresponding to each Parikh vector of the language. An example of slender commutative language is $L = 1^+$, in this case, $d_L(m) = 1$ for every $m \geqslant 1$. For a polynomial commutative language of order n, consider $L = c(1^n0^*)$. It has density $d_L(m) = \binom{m}{n}$ for $m \geqslant n$, that is $\Theta(m^n)$. Finally, if we take $L = (X^n)^+$, it is easy to see that $d_L(m) = l^m$ if m is a multiple of n, and $d_L(m) = 0$ otherwise. When L is regular, the decidability follows by the same argument of the proof of statement (3) in proposition 4.1.1. □

The densities of regular languages in $Sat(\delta_k), Sat(\rho_k), Sat(\gamma_{i,j})$, and $Sat(\varkappa)$ are computable. In the first three cases it is obvious from the expressions given in proposition 4.1.1. It is known that the Parikh set of a regular language can be effectively obtained (cf. [41, 3, 50]), therefore, the density of a regular language in $Sat(\varkappa)$ can also be calculated by using the expression given in the proposition 4.1.2.

4.2 Measures of accuracy of rough approximations of languages

In this section we introduce two definitions of the accuracy of upper rough approximations and establish some of their basic properties. We consider just upper rough approximations because quite often languages do not contain whole classes of the given congruence, thus giving finite lower rough approximations. For example, as it can be directly deduced from our results about density in the previous section, if a language has polynomial density, all its lower $\delta_k, \rho_k, \gamma_{i,j}-$approximations are finite. This may happen even if the language has exponential density. Consider for example, the language $R = X^+u$ where $u \in X^k$ and $k \geqslant 1$. This language is very close to X^*u, its upper δ_k-approximation, but $R_{\delta_k} = \emptyset$.

Some ways to deal with the inexactness of language approximations have been proposed ([26, 19, 8]). In rough set theory [35], the accuracy of the $\theta-$approximation of a set R is expressed as the quotient $|R_\theta|/|R^\theta|$. This notion is not useful for us because of the cases of empty or finite lower approximations noted above and because languages and their approximations are usually infinite sets.

Let R_1 and R_2 be two languages over some alphabet X, such that $R_1 \subseteq R_2$. We use the notion of *natural relative density* given by Berstel in [4],

$$D(R_1, R_2) := \lim_{m\to\infty} |R_1 \cap X^{\leqslant m}|/|R_2 \cap X^{\leqslant m}|$$

whenever this limit exists. The question whether, for a given pair of languages, the first one has a relative density in the second one is decidable for regular languages (see [21]). Berstel proved that this number, even when it exists, is not always rational. Our definitions of accuracy are reformulations of Berstel's definition of relative density.

If $\theta \in \mathrm{Eq}(X^+)$ and $m \geqslant 1$, the *m-accuracy* of the upper $\theta-$approximation of a language $R \subseteq X^+$ is defined by

$$Acc(R, \theta, m) := |R \cap X^{\leqslant m}|/|(R^\theta \cap X^{\leqslant m}|$$

when $R^\theta \cap X^{\leqslant m} \neq \emptyset$, and $Acc(R, \theta, m) := 1$ otherwise.

Observe that since $|R \cap X^{\leqslant m}| = d_R(1) + ... + d_R(m)$, the m-accuracy can be rewritten as $Acc(R, \theta, m) = (d_R(1) + ... + d_R(m))/(d_{R^\theta}(1) + ... + d_{R^\theta}(m))$.

We obtain the following facts immediately from the definition.

4.2.1 Proposition *Let R be a language over an alphabet X and let $\theta, \rho \in \mathrm{Eq}(X^+)$.*

(1) $0 \leq Acc(R, \theta, m) \leq 1$ *for every* $m \geq 1$.

(2) *If* $\theta \subseteq \rho$ *then* $Acc(R, \theta, m) \geq Acc(R, \rho, m)$ *for every* $m \geq 1$.

(3) $Acc(R, \theta, m) = 1$ *for every* $m \geq 1$ *if and only if* $R \in Sat(\theta)$.

4.2.2 Lemma *Let $\theta \in \mathrm{Eq}(X^+)$ and suppose there exists $n \geqslant 1$ such that $w/\theta = \{w\}$ for every $w \in X^{\leq n}$. Then for any $R \subseteq X^+$ and $m \leq n$, $Acc(R, \theta, m) = 1$.*

Proof. It is clear that if $w/\theta = \{w\}$ for every $w \in X^{\leq n}$ then $R^\theta \cap X^{\leqslant m} = R \cap X^{\leqslant m}$ for every $m \leqslant n$. □

The congruences that define the families of $k-$definite, reverse $k-$definite, generalized $i, j-$definite and $k-$testable languages, satisfy the hypothesis of the last lemma.

4.2.3 Corollary *For every regular language $R \subseteq X^+$ and any $m \geqslant 1$, there exist numbers $k \geqslant 1$, $i + j \geqslant 1$ such that $Acc(R, \delta_k, m) = Acc(R, \rho_k, m) = Acc(R, \lambda_k, m) = Acc(R, \gamma_{i,j}, m) = 1$.*

Note that for any language R and any $\theta_k \in \{\delta_k, \rho_k, \lambda_k\}$ with $k \geqslant 0$ we have a non-increasing chain $R^{\theta_0} \supseteq R^{\theta_1} \supseteq \ldots$ of approximations, and according to proposition 4.2.1 and corollary 4.2.3, for every $m \geqslant 0$, a corresponding non-decreasing sequence

$$0 \leqslant Acc(R, \theta_0, m) \leqslant Acc(R, \theta_1, m) \leqslant \ldots \leqslant Acc(R, \theta_k, m) \ldots$$

where $Acc(R, \theta_k, m) = 1$ for every $k > m$. The case of $\gamma_{i,j}$ is similar.

The next proposition states that it is possible to reach values of $m-$accuracy that are arbitrarily close to any given value between 0 and 1 for the cases of $\rho_k, \delta_k, \lambda_k,$ and $\gamma_{i,j}$ approximations.

4.2.4 Proposition *Let X be an alphabet with at least two letters and let $\theta \in \{\rho_k, \delta_k, \lambda_k,$ where $k \geqslant 1$, $i + j \geqslant 1$. For any numbers $q, r \in (0, 1)$ such that $q < r$, there exists a number $m \geqslant 1$ and a regular language $R \subseteq X^+$ such that $q < Acc(R, \theta, m) < r$.*

Proof. Suppose that $L = \{w \in X^+ \mid w/\theta = \{w\}\}$. By the properties of the congruences considered, the set L is finite. Suppose that $|L| = l$. Let us take the language L'. For $m > max\{\lg(w) \mid w \in L\}$, consider a language R over the same alphabet chosen to have the properties $R \cap X^{>m} = L' \cap X^{>m}$, $R \cap X^{\leqslant m} \subseteq L' \cap X^{\leqslant m}$ and suppose that $|R \cap X^{\leqslant m}| = c$, where c is in $\{1, ..., |X^{\leqslant m}| - l\}$. Any language with these properties satisfies $Acc(R, \theta, m) = |R \cap X^{\leqslant m}|/|R^\theta \cap X^{\leqslant m}| = c/|L' \cap X^{\leqslant m}| = c/(|X^{\leqslant m}| - l)$. The number $c/(|X^{\leqslant m}| - l)$ can be fit in (q, r) by choosing values for m and c as follows. Let us call $N = |X^{\leqslant m}| - l$. If we choose m big enough to ensure that $1/N < (r - q)$ then there exists a number c in $\{1, ..., N\}$ such that $q < c/N < r$. □

It is also possible to obtain values of $Acc(R, \varkappa, m)$ arbitrary close to 1.

4.2.5 Proposition *Let X be an at least binary alphabet. For any given $q \in [0, 1)$, there exist a finite language R and $m > 0$ such that $q < Acc(R, \varkappa, m) < 1$.*

Proof. Let us take X such that $|X| = l \geqslant 2$, $x \in X$ and $n \geqslant 2$. The set $F_n = \{w \in X^n \mid |w|_x = 2\}$ satisfies $|F_n| = \binom{n}{2}(l-1)^{n-2} = \frac{n(n-1)}{2}(l-1)^{n-2}$. Now, if R is a subset of F_n such that $\psi(R) = \psi(F_n)$, then $R^\varkappa = F_n$ and $Acc(R, \varkappa, m) = \frac{|R|}{|F_n|}$ for any $m > n$. It is easy to see that the number $|F_n|$ can be made as big as desired, and the number $|R|$ as close to $|F_n|$ as needed. Thus, to obtain the result, it is enough to adjust the number n and the number of words in R, to reach a value such that $q < Acc(R, \varkappa, m) < 1$ holds.

□

Next, we study the asymptotic behavior of the $m-$accuracy when m approaches infinity.

The limit $\lim\limits_{m \to \infty} Acc(R, \theta, m)$ does not always exist for a regular language $R \subseteq X^+$.

4.2.6 Example Consider the δ_2-approximation of the language $R = \{w \in$\{0,1\}$^+ \mid$

$\lg(w) \equiv_2 0\}$. The limit $\lim\limits_{m\to\infty} Acc(R, \delta_2, m)$ does not exist because for even m

$$Acc(R, \delta_2, m) = \sum_{i=1}^{m/2} 2^{2i} / \sum_{i=2}^{m} 2^i = \frac{(2^{m+2} - 4)/3}{2^{m+1} - 4}$$

and for odd m

$$Acc(R, \delta_2, m) = \sum_{i=1}^{(m-1)/2} 2^{2i} / \sum_{i=2}^{m} 2^i = \frac{(2^{m+1} - 4)/3}{2^{m+1} - 4}.$$

Thus the subsequences $\{Acc(R, \delta_2, 2i)\}_{i\geqslant 1}$ and $\{Acc(R, \delta_2, 2i+1)\}_{i\geqslant 1}$ tend to $2/3$ and $1/3$ respectively.

When the limit $\lim\limits_{m\to\infty} Acc(R, \theta, m)$ exists it is denoted $Acc(R, \theta)$, and it is the *accuracy* of the upper $\theta-$approximation of R. The proximity of a language $R \subseteq X^+$ to the approximation R^θ can be described by the number $Acc(R, \theta)$ when it exists.

The following results are direct consequences of the definition of accuracy.

4.2.7 Proposition *Let $R \subseteq X^+$ and $\theta, \rho \in \mathrm{Eq}(X^+)$. If $Acc(R, \theta)$ and $Acc(R, \rho)$ are defined, then*

(1) $0 \leq Acc(R, \theta) \leq 1$.

(2) $Acc(R, \theta) \geq Acc(R, \rho)$ *when* $\theta \subseteq \rho$.

(3) *If R_1 is an infinite language that differs from R just in a finite number of words, then $Acc(R, \theta) = Acc(R_1, \theta)$.*

4.2.8 Corollary *Let us consider a family $\{\theta_k\}_{k\geqslant 1} \subseteq \mathrm{Eq}(X^+)$ that satisfies $\theta_k \supseteq \theta_{k+1}$ for every $k \geqslant 1$.*

(1) *If $Acc(R, \theta_k)$ is defined for every k, then $Acc(R, \theta_1) \leqslant Acc(R, \theta_2) \leqslant \ldots$.*

(2) *If $Acc(R, \theta_k) = 1$ for some $k \geqslant 1$, then $Acc(R, \theta_l) = 1$ for every $l \geqslant k$.*

(3) *If $Acc(R, \theta_k) = 0$ for some $k \geqslant 1$, then $Acc(R, \theta_l) = 0$ for every $l \leqslant k$.*

4.2.9 Lemma *If R is a finite language and $\theta \in \{\rho_k, \delta_k, \lambda_k, \gamma_{i,j}\}$, where $k \geqslant 1$ and $i + j \geqslant 1$, then $Acc(R, \theta) \in \{0, 1\}$. If R is a cofinite language then $Acc(R, \theta) = 1$.*

Proof. Observe that if R is finite, R^θ is either R or an infinite set, depending on the length of the words in R. Therefore, $Acc(R, \theta) = 1$ or 0, respectively. If $R = X^+ \setminus F$ where F is a finite subset of X^+, we have that $R^\theta = X^+ - \{w \in F \mid w/\theta = \{w\}\}$. Consequently, R^θ is either R itself or it differs from R just by a finite number of words, then $Acc(R, \theta) = 1$. □

4.2.10 Lemma *If $R \subseteq X^+$ is a non-empty, non-commutative finite language, then $Acc(R, \varkappa) \in (0, 1)$. If R is any co-finite language, then $Acc(R, \varkappa) = 1$.*

Proof. If R is finite, then $R^\varkappa$ is also finite. Suppose that $|R| = j$ and $|R^\varkappa| = k$ for some $j, k \geqslant 1$. Obviously $j < k$, and then $0 < Acc(R, \varkappa) = j/k < 1$. Suppose now that $R = X^+ \setminus F$, where F is a finite language. It is enough to observe that $R^\varkappa = X^+$, if $F \subseteq R^\varkappa$, $R^\varkappa = R$ if $F \cap R^\varkappa = \emptyset$, and $R^\varkappa = X^+ \setminus F_1$, where $F_1 = F \setminus (F \cap R^\varkappa)$, if $F \cap R^\varkappa \neq \emptyset$ and $F \nsubseteq R^\varkappa$. □

We will often make use of the following fact observed by Berstel in [4] that is a direct consequence of well-known properties of series (see for example Spivak [44], Chapter 22).

4.2.11 Lemma *Let R and S be languages over an alphabet X such that $R \subseteq S$. If the limit $\lim\limits_{m\to\infty} \dfrac{d_R(m)}{d_S(m)}$ exists and the sequence $d_S(m)$ diverges then*

$$\lim_{m\to\infty} \frac{d_R(m)}{d_S(m)} = \lim_{m\to\infty} \frac{d_R(1) + ... + d_R(m)}{d_S(1) + ... + d_S(m)}$$

As a consequence of this result we have

4.2.12 Lemma *Let R be a language over the alphabet X and $\theta \in \mathrm{Eq}(X^+)$ such that the limit $\lim\limits_{m\to\infty} \dfrac{d_R(m)}{d_{R^\theta}(m)}$ exists. Then $Acc(R, \theta)$ exists and equals $\lim\limits_{m\to\infty} \dfrac{d_R(m)}{d_{R^\theta}(m)}$.*

Proof. If R^θ has constant density and $\lim_{m\to\infty} \frac{d_R(m)}{d_{R^\theta}(m)}$ exists, then both $d_R(m)$ and $d_{R^\theta}(m)$ have a constant value from some $N \geqslant 1$ on, and thus the result is clearly true. If R^θ has at least polynomial density, then $d_{R^\theta}(m)$ diverges and by hypothesis $\lim_{m\to\infty} \frac{d_R(m)}{d_{R^\theta}(m)}$ exists. Thus, we can apply the previous result to obtain

$$Acc(R,\theta) = \lim_{m\to\infty} \frac{d_R(1) + ... + d_R(m)}{d_{R^\theta}(1) + ... + d_{R^\theta}(m)} = \lim_{m\to\infty} \frac{d_R(m)}{d_{R^\theta}(m)}$$

□

Observe that under the hypothesis of the last lemma

$$Acc(R,\theta) = \lim_{m\to\infty} \frac{d_R(m)}{d_{R^\theta}(m)} = \lim_{m\to\infty} \frac{d_R(m)}{d_R(m) + d_{R^\theta - R}(m)} = \frac{1}{1 + \lim\limits_{m\to\infty} \frac{d_{R^\theta - R}(m)}{d_R(m)}}$$

Let us define, whenever it exists,

$$T(R,\theta) = \lim_{m\to\infty} \frac{d_{(R^\theta - R)}(m)}{d_R(m)}$$

This number gives an idea of the size of what is added to R to obtain the approximation R^θ.

4.2.13 Lemma *Let us consider a congruence $\theta \in \mathrm{Eq}(X^+)$ and a language R over X such that the limit $\lim_{m\to\infty} \frac{d_R(m)}{d_{R^\theta}(m)}$ exists. Then $Acc(R,\theta) = 1$ if and only if $T(R,\theta) = 0$.*

In the next two sections we compute the accuracy of the upper approximations of languages of different densities in the families under study.

4.3 Languages over a one-letter alphabet

Let us consider the case of languages over a one-letter alphabet $X = \{0\}$. It is well known and easy to see that if R is an infinite one-letter regular language, $R =$

$0^{x_1} + ... + 0^{x_l} + 0^p(0^r)^*(0^{y_1} + ... + 0^{y_s})$ for some numbers $l, x_1, ..., x_l, r, s, y_1, ..., y_s$ such that $l \geqslant 0$, $0 \leqslant x_1 < ... < x_l \leqslant p$, $r \geqslant 1$, $s \geqslant 0$ and $0 \leqslant y_1 < ... < y_s < r$. The statement (3) in proposition 4.2.7 allows us to assume w.l.g. that $x_i = 0$ for every $i \in \{1, ..., l\}$ and that

(1) $R = (0^r)^*(0^{c_1} + ... + 0^{c_s})$, where $s \geqslant 1$ and $0 \leqslant c_1 < ... < c_s < r$.

We can assume that $s \geqslant 1$, because if $s = 0$ the same results can be straightforwardly obtained. Observe that over a unary alphabet, $\delta_k = \rho_k = \lambda_k = \gamma_{i,j}$, where $k = max\{i, j\} \geqslant 1$. According to this, all the $\{\rho_k, \delta_k, \lambda_k, \gamma_{i,j}\}-$approximations can be handled together for the case of a one-letter alphabet. In this section we refer to any of those congruences as θ_k. Their classes are $u/\theta_k = 0^k0^*$ if $u \in X^{\geqslant k}$ and $u/\theta_k = \{u\}$ if $u \in X^{<k}$. In what follows, R is a language as in (1).

4.3.1 Lemma $|R \cap X^{\leqslant nr}| = ns$ *for every* $n \geqslant 1$.

4.3.2 Lemma *If* R *has at least one word* $u \in X^{\geqslant k}$, *then* $R^{\theta_k} = 0^k0^* \cup (R \cap X^{<k})$ *and* $|R^{\theta_k} \cap X^{\leqslant m}| = m - k + 1 + |R \cap X^{<k}|$ *for every* $m > k$.

In the next two lemmas $c = |R \cap X^{<k}|$.

4.3.3 Lemma $Acc(R, \theta_k, nr) = ns/(nr - k + 1 + c)$ *where* n *is such that* $nr > k$.

4.3.4 Lemma $\lfloor m/r \rfloor s/(m-k+1+c) \leqslant Acc(R, \theta_k, m) \leqslant (\lfloor m/r \rfloor + 1)s/(m-k+1+c)$ *for every* $m > k$.

4.3.5 Proposition $Acc(R, \theta_k) = s/r$ *for every* $k \geqslant 1$.

Proof. As both the upper and the lower bound given in the previous lemma for $Acc(R, \theta_k, m)$ tend to s/r when m goes to infinity, $Acc(R, \theta_k, m)$ does, by the Sandwich Theorem. □

According to this result the accuracy of approximations in the case of a one-letter regular language always exists and it is a rational number, but it cannot be improved by taking larger $k's$. Observe also that the accuracy in this case can be effectively computed.

4.3.6 Corollary *For every $k \geqslant 1$ and numbers n and m such that $0 < m < n$, there exists an infinite one-letter regular language R such that $Acc(R, \theta_k) = m/n$.*

4.3.7 Example Consider the θ_k-approximation of $R = (0^n)^+$ where $k, n \geqslant 1$. For $m \geq k, n$

$$Acc(R, \theta_k, m) = |\{0^n, 0^{2n}, ..., 0^{\lfloor m/n \rfloor n}\}|/|\{0^k, 0^k 0, ..., 0^m\}| = \lfloor m/n \rfloor / (m - k + 1)$$

It is not hard to see that the limit of $Acc(R, \theta_k, m)$ when m aproaches infinity is $1/n$.

4.4 Languages over an at least binary alphabet

Throughout this section, X is an at least binary alphabet. We start by giving an example.

4.4.1 Example Let us consider the language $R = X^*0 \setminus 0^+0$, where $X = \{0, 1\}$, and its 1-definite approximation. Since $R \cap X^m = X^{m-1}0 \setminus \{0^m\}$ and $R^{\delta_1} \cap X^m = X^{m-1}0$, we obtain $Acc(R, \delta_1, m+1) = \frac{(2^m - 1)}{2^m}$. Therefore, $Acc(R, \delta_1)$ exists and equals 1.

Observe that this language is "almost" $1-$definite. Only one word of each length has been removed from X^*0, which is the $1-$definite approximation of R. It is a particular case of the following general fact.

4.4.2 Lemma *Let R be an exponential density regular language over X, $\theta \in \mathrm{Eq}(X^+)$ and suppose that the limit $\lim\limits_{m \to \infty} \dfrac{d_R(m)}{d_{R^\theta}(m)}$ exists. If $R^\theta \setminus R$ has polynomial density, then $Acc(R, \theta) = 1$.*

Proof. If $R^\theta \setminus R$ has polynomial density and R has exponential density then $T(R, \theta) = 0$. Therefore, according to lemma 4.2.13 $Acc(R, \theta) = 1$. □

The value of the accuracy cannot be predicted if the difference between the language and its approximation has exponential density. According to the results of Szilard, Yu and Zhang in [47], an exponential density language can only have a $2^{\Theta(cm)}$

density function, where c is a constant. This means that $T(R,\theta) = \lim\limits_{m\to\infty} \dfrac{2^{c_1 m + c_2}}{2^{b_1 m + b_2}}$, where c_1, c_2, b_1, b_2 are constants. Therefore the possible values of $T(R,\theta)$ are $\infty, 0$ or some finite positive constant, when $c_1 > b_1$, $c_1 < b_1$ or $c_1 = b_1$, respectively. Accordingly, the accuracy $Acc(R,\theta)$ can be $0, 1$ or $0 < Acc(R,\theta) < 1$.

The following result is a consequence of proposition 4.1.1.

4.4.3 Corollary *For every regular language $R \subseteq X^+$ of at most polynomial density and any $\theta \in \{\delta_k, \rho_k, \gamma_{i,j}\}$, $Acc(R,\theta) = 0$.*

4.4.4 Example Let us consider the 2−definite approximation of the language $R = 01^*0$ over the alphabet $X = \{0,1\}$. Since $R^{\delta_2} = X^*10 \cup X^*00$, we obtain for $m \geq 2$, $Acc(R,\delta_2,m) = 1/2^{m-1}$. Therefore, the limit of $Acc(R,\delta_2,m)$ exists and equals 0.

Unlike the case of the families in the hypothesis of this corollary, polynomial languages may have non-zero accuracy values when they are approximated by members of the commutative family, as the following proposition shows.

4.4.5 Proposition *For every $n \geqslant 2$, there exists a polynomial language $R \subseteq \{0,1\}^+$ such that $Acc(R,\varkappa) = 1/n$.*

Proof. Let us define $R = 0^{n-1}(0^n)^*1(0^n)^*$, $n \geqslant 2$. It has density $d_R(m) = k$ if $m = kn$ for some $k \geqslant 1$, and $d_R(m) = 0$ otherwise. Its commutative approximation is $R^{\varkappa} = 0^*10^* \cap (X^n)^+$, that has density $d_{R^{\varkappa}}(m) = m$ if $m = kn$ for some $k \geqslant 1$, and $d_{R^{\varkappa}}(m) = 0$ otherwise. Therefore, $Acc(R,\varkappa) = \lim_{m\to\infty} \frac{\sum_{i=1}^{m} d_R(i)}{\sum_{i=1}^{m} d_{R^{\varkappa}}(i)} = \lim_{k\to\infty} \frac{1+2+\ldots+k}{n+2n+\ldots+kn} = 1/n$. □

As a consequence of corollary 4.4.3, only exponential density languages can be expected to have nonzero accuracy values for the k-definite, reverse k-definite and generalized i,j-definite cases. Let us consider for example the language $R = \{0,1\}^*0^{n+1}$, where $n \geqslant 1$, and its 1−definite approximation. We obtain for $m > n$, $Acc(R,\delta_1,m) = 2^{m-n-1}/2^{m-1}$ and thus $Acc(R,\delta_1) = 1/2^n$.

4.4.6 Proposition *For every number $q \in [0,1)$ and $\theta \in \{\delta_k, \rho_k, \lambda_k, \gamma_{i,j}\}$, where $k \geqslant 1$ and $i+j \geqslant 1$, there exists an alphabet X and an exponential density regular language R over X such that $q < Acc(R,\theta) < 1$.*

Proof. Let us consider the δ_k-approximation of the language $R = u_1X^* \cup ... \cup u_rX^*$ over an alphabet X, where $n > k$, $1 \leqslant r \leqslant |X|^n$ and $u_1, ..., u_r \in X^n$. As the δ_k-approximation of R is $X^{\geqslant k}$ we obtain $Acc(R, \delta_k) = \lim_{m\to\infty} r|X|^{m-n}/|X|^m = r/|X|^n$. Thus, to satisfy our thesis it suffices to choose $|X|$, r and n such that $1/|X|^n < 1-q$ and $r = |X|^n - 1$. The cases of ρ_k and $\gamma_{i,j}$ are similar. For the case of λ_k let us take an alphabet X such that $0 \notin X$ and the language $R = 0^lX^* \cup 0^{l-1}X^* \cup ... \cup 0^kX^*$ where $k < l$. Then

$$Acc(R, \lambda_k) = \lim_{m\to\infty} \frac{|X|^{m-l} + |X|^{m-l+1} + ... + |X|^{m-k}}{|X|^{m-k} + |X|^{m-k-1} + ... + 1} = 1 - \frac{1}{|X|^{l-k+1}}$$

and we can take l and $|X|$ such that $1/|X|^{l-k+1} < 1-q$. □

Because of the fact that $\delta_k, \rho_k, \gamma_{i,j}$ form descending chains for increasing k, i, j, better approximations of an exponential language can be obtained by taking greater elements of that chains, as the following example shows.

4.4.7 Example Consider the language $R = 0^+ + 1X^*$, and $k > 1$, then $d_R(m) = 2^{m-1} + 1$, $R^{\rho_k} = 0 + 0^2 + \cdots + 0^{k-1} + 0^kX^* + 1X^*$ and $d_{R^{\rho_k}}(m) = 2^{m-k} + 2^{m-1}$ for $m \geqslant k$. Therefore,

$$Acc(R, \rho_k) = \lim_{m\to\infty} \frac{2^{m-1}+1}{2^{m-k}+2^{m-1}} = \lim_{m\to\infty} \frac{2^{m-1}}{2^{m-1}(2^{-k+1}+1)} = \frac{1}{1+\frac{1}{2^{k-1}}}$$

So that when we take bigger k's, the accuracy improves.

Next, we show some examples of the possible situations that may appear when an exponential language is approximated by a member of the commutative family. The densities of the $\varkappa$-approximations are calculated by using the expression obtained in proposition 4.1.2.

4.4.8 Example Let $X = \{0, 1\}$.

(1) For the language $R = (00+11)^+$ we have

$Acc(R, \varkappa) = \lim_{2m\to\infty} \frac{2+2^2+\ldots+2^m}{2+2^3+\ldots 2^{2m-1}} = \lim_{m\to\infty} 2(\frac{1}{2})^m = 0.$

(2) Let $R = (00+01+10)^+$. Despite the fact that this language has many more words of even length than the language in (1), it is not difficult to see that the accuracy is again 0, $Acc(R, \varkappa) = \lim_{m\to\infty} (\frac{3}{4})^m = 0$.

(3) Finally, to see some non-trivial exponential language that is well approximated by a commutative language, let us consider $R = \{w \in (X^2)^+ \mid 1^n \notin \mathrm{sw}_n(w), n \geqslant 2\}$. The densities are $d_R(2m) = ((2^{2m} + \binom{2m}{m})/2) - (m-1)(2m-1)$ and $d_{R^{\varkappa}}(2m) = (2^{2m} + \binom{2m}{m})/2$. Then we obtain $Acc(R, \varkappa) = 1$.

4.5 Concluding remarks

In this chapter several facts have been established that describe the behavior of the density of the approximations with respect to the density of the object language. In particular, we focused on the accuracy of the upper rough approximation of a regular language by a language in the k-definite, reverse k-definite, i, j-definite, k-testable and commutative families. We showed the asymptotic behavior of the relative density for a one-letter alphabet and for the general case of an arbitrary alphabet, and found the attainable values of accuracy in each case.

Chapter 5

Comparisons with previous work

In the paper [33] Păun, Polkowski and Skowron introduce several indiscernibility relations among strings that are equivalence or tolerance relations, and study lower and upper rough approximations of languages defined by them. After presenting some general facts about the relations, they consider the problem of approximating context-free languages and find that in most cases the obtained approximations are regular. Finally, they suggest some possible variants of the relations. In this chapter we study some of these indiscernibility relations from our point of view, their connections with the families we already considered, and how some of our results can be applied to them. Firstly, we consider the relations as they were defined in [33], and after that we introduce modifications. The original notation is simplified to conform with ours. For example, the upper rough approximation of a language L under the relation $Pref_k$ is denoted $\overline{Pref_k(L)}$ in [33], but here by L^{P_k}. We also restrict the relations defined on X^* to X^+ to fit in our general framework. Each relation $R(X)$ on X^+ considered in this chapter, will be written R when there is no need to specify the alphabet.

5.1 Working with the original relations

The relations P_k, A_k and M_k studied in this section are those named $Pref_k, ASub_k$ and $MSub_k$ in [33]. The relation S_k was not defined there, but we included it as it seems natural to consider suffixes along with prefixes.

5.1.1 The relations $P_k(X)$ and $S_k(X)$

The relation $P_k(X)$ on X^+, $k \geqslant 1$, is defined by

$$P_k(X) := \{(u, v) \mid u, v \in X^+, \mathrm{lg}(u) = \mathrm{lg}(v), \mathrm{pref}_k(u) = \mathrm{pref}_k(v)\}.$$

The condition $\mathrm{lg}(u) = \mathrm{lg}(v)$ will be called the *length condition.*

$P_k(X)$ is in $\mathrm{Con}(X^+)$ and it has infinite index because of the length condition, but the family $\{P_k(X)\}_X$ is not consistent. Indeed, if $\varphi : X^+ \to Y^+$ is a morphism that is not length-preserving, then $P_k(X) \nsubseteq \varphi \circ P_k(Y) \circ \varphi^{-1}$ because the fact that for two words $u, v \in X^+$, $\mathrm{lg}(u) = \mathrm{lg}(v)$, does not imply that $\mathrm{lg}(u\varphi) = \mathrm{lg}(v\varphi)$. For example, if $X = \{a, b\}$, $Y = \{c\}$ and $\varphi : X^+ \to Y^+$ is defined by $a\varphi = c$, $b\varphi = cc$, then $(aa, ab) \in P_1(X)$ but $(aa\varphi, ab\varphi) \notin P_1(Y)$ because $aa\varphi = cc$ and $ab\varphi = ccc$ have different lengths. If we drop the length condition in the definition of $P_k(X)$, it becomes $\rho_k(X)$ and we would have the associated family of reverse k-definite languages.

If the congruences of a family include the length condition, they need to be finer than $\varkappa(X)$, otherwise we can always define a morphism that is not length-preserving, and does not preserve the congruence, as in the above example.

5.1.1 Proposition *For every $k \geqslant 1$,*

(1) $X^+/P_k = \{\{u\} \mid u \in X^{<k}\} \cup \{uX^j \mid u \in X^k, j \geqslant 0\}$,

(2) *the P_k-approximations are finite unions of languages of the form $uX^* \cap \bigcup_{j\in I} X^j$ where $u \in X^{\leqslant k}$, $I \subseteq \mathbb{N}$, and*

(3) *if $L \subseteq X^+$ is a context-free language, then the P_k-approximations of L are finite unions of languages of the form $uX^* \cap \bigcup_{j\in I} X^j$, where $u \in X^{\leqslant k}$ and $I \subseteq \mathbb{N}$ forms an arithmetic sequence.*

Proof. The first statement is obvious from the definition of P_k. To prove (2), it is enough to observe that for every language L and $u \in X^k$, the set $uX^* \cap \bigcup_{j\in I} X^{k+j}$ where $I = \{\mathrm{lg}(v) \mid v \in u^{-1}L\}$ is added to the approximation, and if $u \in X^{<k} \cap L$, then $\{u\}$ is added. Finally, for any vector $\overline{v} = (s_1, ..., s_n)$, $n = |X|$ and $s_1, ..., s_n \geqslant 0$,

we denote $|\overline{v}| = s_1 + ... + s_n$. If L is context-free, its Parikh set is semilinear. As $u^{-1}L$ is also context-free for every $u \in X^+$, its Parikh set is a finite union of sets of the form $\{\overline{v_0} + t_1\overline{v_1} + ... + t_m\overline{v_m} \mid t_1, ..., t_m \geqslant 0\}$ where $\overline{v_0}, ..., \overline{v_m} \in \mathbb{N}^n$, and thus, the set of possible lengths of words in $u^{-1}L$ is a finite union of sets of the form $J = \{|\overline{v_0}| + t_1|\overline{v_1}| + ... + t_m|\overline{v_m}| \mid t_1, ..., t_m \geqslant 0\}$. If $\gcd(|\overline{v_1}|, ..., |\overline{v_m}|) = 1$, then $J = \{s \mid s \geqslant N\} \cup F$ for some $N \geqslant 0$ and F a finite subset of $\mathbb{N}$, otherwise J is of the form $\{n_0 + n_1 j \mid j \geq 0\}$ for some $n_0 \geqslant 0$ and $n_1 > 1$. It remains to be observed that the sets I are finite unions of the sets J associated with each prefix of L of length k. □

5.1.2 Example If $L = \{a^i b^i \mid i \geqslant 1\}$, then $L^{P_3} = \bigcup_{i \geqslant 1} aaaX^{2i+1} \cup aabX \cup \{ab\} = (aaaX^* \cap \bigcup_{i \geqslant 3} X^{2i}) \cup (aabX^* \cap X^4) \cup \{ab\}$.

Note that for every language L, L^{ρ_k} is regular, in fact reverse k-definite. In [33] it is proved that L^{P_k} is regular for any context-free language L, but this does not hold for all languages. Take for example $X = \{a\}$ and $L = \{a^p \mid p \in \mathbb{P}\}$. In this case $L^{P_k} = L$ for every k.

As $P_k \subseteq \rho_k$, we have $L^{P_k} \subseteq L^{\rho_k}$ and $L_{P_k} \supseteq L_{\rho_k}$, but even for regular languages, it is not true in general that $L^{P_k} = L^{\rho_k}$ or $L_{P_k} = L_{\rho_k}$. For example, if L is $(aaa)^+$, then $L^{P_3} = \bigcup_{n \geqslant 0} aaaX^{3n}$ and $L_{P_3} = \{aaa\}$ while $L^{\rho_3} = aaaX^*$ and $L_{\rho_3} = \emptyset$.

5.1.3 Proposition *Let $L \subseteq X^+$ and $k \geqslant 1$.*

(1) *If L is a finite language, then $L^{P_k} = L^{\rho_k}$ iff $L \subseteq X^{<k}$.*

(2) *If L is an infinite language, then $L^{P_k} = L^{\rho_k}$ iff $d_{u^{-1}L}(n) \neq 0$ for every $u \in pref_k(L) \cap X^k$ and $n \geqslant 0$.*

(3) *If L is a finite language, then $L_{P_k} = L_{\rho_k}$ iff for every $u \in pref_k(L) \cap X^k$ and $j \geqslant 0$, $uX^j \nsubseteq L$.*

(4) *If L is an infinite language, then $L_{P_k} = L_{\rho_k}$ iff $uX^* \subseteq L$ for each $u \in pref_k(L) \cap X^k$ such that $uX^j \subseteq L$ for some $j \geqslant 0$.*

Proof. The fact (1) is obvious. To prove (2) it is enough to observe that $L^{P_k} = L^{\rho_k}$ iff for each k-prefix of length k of a word in L, the language contains words of every length with that prefix. But this is exactly the condition that $d_{u^{-1}L}(n) \neq 0$ for every $u \in pref_k(L) \cap X^k$ and $n \geqslant 0$. The claims (3) and (4) follow from the definitions similarly as (1) and (2). □

Let $S_L := \sum_{i=0}^{\infty} s_i x^i$ be the formal power series where $s_i := d_L(i)$, that is to say, the i^{th} coefficient of the series is the number of words of length i in the language L. For a context-free language, the series S_L can be obtained by using the procedure known as Schützenberger's method (cf. [9]).

Let $G = (V, X, P, S)$ be an unambiguous context-free grammar, where V denotes the set of non-terminals, X the set of terminals, P the set of productions, S is the initial symbol, and let L_G be the generated language. The sets $(V \cup X)^+$ and $(V(x) \cup \{x\})^+$, where $V(x) = \{A(x) \mid A \in V\}$, are viewed as commutative semigroups, and the morphism $\Theta : (V \cup X)^+ \to (V(x) \cup \{x\})^+$ is defined as follows: $\Theta(a) = x$ for every $a \in X$, and $\Theta(A) = A(x)$ for every $A \in V$. For example, the words Aaa, aAa, aaA in $(V \cup X)^+$ are regarded as the same element, and they are sent by Θ to $A(x)x^2 (= xA(x)x = x^2A(x))$ in $(V(x) \cup \{x\})^+$.

We then associate with every set of productions $A \to e_1, A \to e_2, ..., A \to e_n$ in P, where $A \in V$ and $e_i \in (V \cup X)^*$, the algebraic equation

$$\Theta(A) = \Theta(e_1) + ... + \Theta(e_n)$$

where $\Theta(e_i) = 1$ if $e_i = \varepsilon$. The resulting system is then solved for $S(x)$ and it gives the generating function of the series S_{L_G}.

5.1.4 Example The grammar $G = \{\{S\}, \{a, b\}, \{S \to aSb|ab\}, S\}$ generates $L = \{a^i b^i \mid i \geqslant 1\}$. Applying the Schützenberger method we obtain $S(x) = S(x)x^2 + x^2$ and then $S(x) = \frac{x^2}{1-x^2}$, whose expansion has the coefficients $0, 0, 1, 0, 1, 0, ...$

When L is an infinite language, the condition to have $L^{P_k} = L^{\rho_k}$, given in Proposition 5.1.3 (2), is that for every $u \in \text{pref}_k(L) \cap X^k$ and $n \geqslant 0$, $d_{u^{-1}L}(n) \neq 0$. We will now outline a procedure to find out whether this condition holds for a regular language L, given by an unambiguous regular grammar. It was proved in [9] that if this is the case, the corresponding generating function is rational. Note that such a grammar can be obtained for any regular language from a deterministic automaton, by including for any transition labeled by a letter a from a state A to a state B,

- the production $A \to aB$ if B is not a final state, and
- the productions $A \to aB$ and $A \to a$, if B is final.

Moreover, include the rule $S \to \varepsilon$ if the initial state S is final.

The procedure is as follows:

(1) Calculate the set $\text{pref}_k(L) \cap X^k = \{u_1, ..., u_n\}$

(2) Form a grammar for $u_i^{-1}L$ for every $i = 1, ..., n$

(3) Apply Schützenberger's method to find the formal power series $\sum_{j \in I_i} s_j x^j$ of $u_i^{-1}L$, where I_i is the index set corresponding to the non-zero coefficients.

(4) $L^{P_k} = (\bigcup_{i=1}^{n} \bigcup_{j \in I_i} u_i X^j) \cup (L \cap X^{<k})$

A grammar for the left derivative of a regular language in step (2) can be easily built, and if G is a regular unambiguous grammar, the whole procedure is effective. In this case, the formal power series in step (3) is rational and thus its coefficients can be efficiently computed, for example by using the method of partial fractions.

5.1.5 Example Let $L = \{u \in \{a,b\}^+ \mid \lg(u) \equiv_3 1\}$. An unambiguous grammar for L is $G = (S, X, V, P)$ where S is the initial symbol, $X = \{a, b\}$ is the set of terminal symbols, $V = \{S, A, B\}$ is the set of non-terminals, and P is the following set of productions: $S \to aA|bA|a|b$, $A \to aB|bB$, $B \to aS|bS$. The set of 2-prefixes of length 2 of L is $\text{pref}_2(L) \cap X^2 = X^2$. A grammar for $u^{-1}L$, where $u \in \text{pref}_2(L)$, is obtained just by considering B as the new initial symbol. Applying Schützenberger's method to this grammar we obtain $B = \frac{4x^2}{1-8x^3}$ that has the sequence of coefficients

$0, 0, 2^2, 0, 0, 2^5, 0, 0, 2^8, ...$ As the sequence has zeros, according to our condition, $L^{\rho_2} \neq L^{P_2}$. In fact, the P_2-approximation turns out to be $L^{P_2} = \{a, b\} \cup X^2(X^2 + X^5 + X^8 + ...) = X + X^4 + X^7 + X^{10} + ... = L$, while $L^{\rho_2} = X^+$.

We present a second example of an application of the procedure, in which the P_k-approximation of the language is not trivial.

5.1.6 Example Let $X = \{a, b\}$ and $L = (aaa)^+$. An unambiguous grammar for L is $G = (S, X, V, P)$ where $V = \{S\}$, and $P = \{S \to aaaS, S \to aaa\}$. The set of 3-prefixes is obviously $\text{pref}_3(L) \cap X^3 = \{aaa\}$. A grammar for $aaa^{-1}L$, is obtained by just replacing $S \to aaa$ by $S \to \varepsilon$. Applying Schützenberger's method to this grammar we obtain $S = \frac{1}{1-x^3}$, which has the sequence of coefficients $1, 0, 0, 1, 0, 0, 1, 0, 0, ...$ As the sequence has zeros, this implies again that $L^{\rho_3} \neq L^{P_3}$, and in this case $L^{P_3} = aaa(\varepsilon + X^3 + X^6 + ...)$ while $L^{\rho_3} = aaaX^*$.

In some cases the process can be carried out also for context-free languages.

5.1.7 Example Let $L \subseteq \{a, b\}^+$ be $\{a^i b^i \mid i \geqslant 1\}$. The set of 2-prefixes of L is $\text{pref}_2(L) = \{ab, aa\}$, and $(ab)^{-1}L = \{\varepsilon\}$, $(aa)^{-1}L = \{a^{i-2}b^i \mid i \geq 2\}$. Using Schützenberger's procedure we obtain $S = 1$ for $(ab)^{-1}L$ and $S = \frac{x^2}{1-x^2}$ for $(aa)^{-1}L$. The sequence of coefficients of the latter is $0, 0, 1, 0, 1, 0, ...$ Thus, we obtain $L^{P_2} = \{ab\} \cup aaX^2 \cup aaX^4 \cup ...$, while $L^{\rho_2} = aaX^* \cup abX^*$.

Let us define the relation $S_k(X)$ over X^+ by

$$S_k(X) := \{(u, v) \mid u, v \in X^+, \text{lg}(u) = \text{lg}(v), \text{suff}_k(u) = \text{suff}_k(v)\}$$

If we denote by u^R the reversal of a word u and L^R the reversal of a language L, then $Lu^{-1} = ((u^{-1})^R L^R)^R$ and all the arguments used for P_k-approximations can easily be modified to apply to $S_k(X)$.

5.1.2 The relations $A_k(X)$ and $M_k(X)$

For $w, u \in X^+$, the *multiplicity* of w in u is the number

$$\mu(w,u) := |\{u_1 \in X^* \mid u = u_1 w u_2, \text{for some } u_2 \in X^*\}|.$$

For any $k \geqslant 1$, the relations $A_k(X)$ and $M_k(X)$ on X^+ are defined as follows:

$$A_k(X) := \{(u,v) \mid u,v \in X^{>k}, \mathrm{lg}(u) = \mathrm{lg}(v), \mathrm{sw}_k(u) = \mathrm{sw}_k(v)\} \\ \cup \{(u,u) \mid u \in X^{\leqslant k}\}.$$

$$M_k(X) := \{(u,v) \mid u,v \in X^{>k}, (\forall w \in X^k)\mu(w,u) = \mu(w,v)\} \\ \cup \{(u,u) \mid u \in X^{\leqslant k}\}.$$

As observed in [33], the length condition makes no difference for M_k, because the length of a word u is completely determined by $\mathrm{sw}_k(u)$ and the multiplicities of the members of that set, so we may drop this condition from the definition of M_k. None of the relations A_k and M_k are congruences on X^+. For example, the words *bab* and *aba* are related by A_2, but if we add the prefix b, we get $(bbab, baba) \notin A_2$. The same example shows that M_2 is not a congruence. Moreover, none of them are of finite index because each class has a finite number of elements. The relation A_k and the congruence λ_k, associated with k-testable languages, are incomparable. For example, $(ababa, aba) \in \lambda_2 \setminus A_2$ and $(aba, bab) \in A_2 \setminus \lambda_2$. The same examples show that M_k and λ_k are also incomparable.

For any word $u \in X^+$ and any $k \geqslant 1$, let us denote the complement of $\mathrm{sw}_k(u)$ within X^k by $\mathrm{nsw}_k(u)$.

5.1.8 Proposition *Let $L \subseteq X^+$, $k \geqslant 1$ and $u \in X^+$.*

(1) *$A_{k+1} \subseteq A_k$, and the inclusion is proper if $|X| > 1$.*

(2) *If* $L \subseteq X^{\leqslant k}$*, then* $L^{A_k} = L_{A_k} = L^{M_k} = L_{M_k} = L$*.*

(3) *If* $\lg(u) > k$*, then* $[u]_{A_k} = (\bigcap_{s\in \mathrm{sw}_k(u)} X^*sX^* \setminus \bigcup_{t\in \mathrm{nsw}_k(u)} X^*tX^*) \cap X^{\lg(u)}$*.*

(4) *The rough approximations of* L *under* A_k *are finite unions of sets of the form* $C \cap (\bigcup_{j\in I} X^j)$ *where* C *is a* k*-testable language and* $I \subseteq \mathbb{N}$*.*

(5) *If* L *is context-free, then the rough approximations under* A_k *are regular and the sets* I *in (4) form arithmetic sequences.*

Proof. The fact that $A_k = \Delta_{X^+}$ for $|X| = 1$ is obvious. The (non-strict) inclusion in (1) was proved in [33]. To see that it is proper if $|X| > 1$, observe that for every $x, y \in X$, the words $x^{k-1}yx^{k-1}$ and $x^{k-2}yx^{k-1}y$, are in $A_k \setminus A_{k+1}$. The fact (2) is obvious because for every word u, if $\lg(u) \leqslant k$, both $[u]_{A_k}$ and $[u]_{M_k}$ are $\{u\}$. The expression in (3) comes directly from the definition of the relation. To see that (4) holds, observe that there is a finite number of sets $\bigcap_{s\in \mathrm{sw}_k(u)} X^*sX^* \setminus \bigcup_{t\in \mathrm{nsw}_k(u)} X^*tX^*$ for $u \in L$, and that they are k-testable languages. As the set $\mathrm{sw}_k(L) = \bigcup_{u\in L} \mathrm{sw}_k(u)$ is finite, let us take $\{u_1, ..., u_n\} \subseteq L$ such that $\mathrm{sw}_k(L) = \bigcup_{i=1}^n \mathrm{sw}_k(u_i)$. If we denote $I_i = \{\lg(u) \mid u \in L, \mathrm{sw}_k(u) = \mathrm{sw}_k(u_i)\}$, and $C_i = \bigcap_{s\in \mathrm{sw}_k(u_i)} X^*sX^* \setminus \bigcup_{t\in \mathrm{nsw}_k(u_i)} X^*tX^*$ for $i = 1, ..., n$, $n \geqslant 1$, then $L^{A_k} = (\bigcup_{i=1}^n C_i \cap \bigcup_{j\in I_i} X^j) \cup (L \cap X^{\leqslant k})$. The lower approximations L_{A_k} are also finite unions of this type of sets. Suppose now, to prove (5), that L is a context-free language. The fact that its rough approximation under A_k is regular was proved in [33] (Proposition 6.4), but we offer an alternative proof. If L is context-free, the sets $C_i \cap L$, are also context-free and as the sequences I_i are the sets of possible lengths of words in $C_i \cap L$, the argument used in the proof of statement (3) of lemma 5.1.1 can be applied to prove that the sets I_i form arithmetic sequences. Now, let us take a regular language R such that it has the same Parikh set as $C_i \cap L$ and define the substitution σ from X to subsets of X^+ by putting $\sigma(a) = X$ for every $a \in X$. It gives $\sigma(R) = \bigcup_{j\in I_i} X^j$, that is a regular set, and therefore each set $C_i \cap \bigcup_{j\in I_i} X^j$ is regular. □

5.1.9 Example If L is the context-free language $\{a^nb^n \mid n \geqslant 1\}$, then $\mathrm{sw}_2(L) =$

$\{aa, ab, bb\}$ and $L^{A_2} = a^2a^*b^*b^2 \cap (X^2)^+ \cup \{ab\} = [(X^*aaX^* \cap X^*abX^* \cap X^*bbX^*) \setminus X^*baX^*] \cap \bigcup_{i \geqslant 2} X^{2i} \cup \{ab\}$.

The approximations under A_k will then correspond to finite unions of "slices" of locally testable languages of the lengths existing in the language. If we drop the length condition in the definition of A_k, the classes become just locally testable languages (as it is clear from (3) in the last proposition), and then the approximations under A_k thus modified, are also regular.

5.2 Working with modified relations

5.2.1 Modifying $A_k(X)$

To get a refinement of $A_k(X)$ that is a congruence on X^+, let us define $\overline{A_k}(X) \subseteq A_k(X)$ by

$$(u, v) \in \overline{A_k}(X) \Leftrightarrow (u, v) \in A_k(X), \mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v), \mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v).$$

This is clearly a refinement of $\lambda_k(X)$. In fact, $(u, v) \in \overline{A_k}(X)$ iff $(u, v) \in \lambda_k(X)$ and $\lg(u) = \lg(v)$. The relation thus modified is a congruence on X^+ of infinite index, but the associated family is not consistent because it does not satisfy $\overline{A_k}(X) \subseteq \varkappa(X)$, for every $k \geqslant 1$.

5.2.1 Proposition *Let $L \subseteq X^+$, $k \geqslant 1$ and $u \in X^+$.*

(1) *If $\lg(u) \leqslant k$, then $[u]_{\overline{A_k}} = \{u\}$ and hence $L^{\overline{A_k}} = L_{\overline{A_k}} = L$ if $L \subseteq X^{\leqslant k}$.*

(2) *If* $\lg(u) > k$, *then*

$$[u]_{\overline{A_k}} = ((\bigcap_{s\in \mathrm{sw}_k(u)} X^*sX^*) \cap \mathrm{pref}_{k-1}(u)X^* \cap X^*\mathrm{suff}_{k-1}(u) \setminus (\bigcup_{t\in \mathrm{nsw}_k(u)} X^*tX^*)) \cap X^{\lg(u)}.$$

(3) *The rough approximations of* L *under* $\overline{A_k}$ *are finite unions of sets of the form* $C \cap \bigcup_{j\in I} X^j$ *where* C *is a* k*-testable language and* $I \subseteq \mathbb{N}$.

(4) *If* L *is context-free, then the sets* I *in (3) form arithmetic sequences and the rough approximations of* L *under* $\overline{A_k}$ *are regular.*

(5) $\overline{A_k} \subset \lambda_k$.

(6) $\overline{A_{k+1}} \subseteq \overline{A_k}$, *and the inclusion is proper if* $|X| > 1$.

(7) $\overline{A_k} \in \mathrm{Con}(X^+) \setminus \mathrm{FCon}(X^+)$.

Proof. Facts (1), (2) and (7) follow directly from the definition of the relation. To prove (3) we take a set of words $\{u_1, ..., u_n\} \subseteq L$ such that for every $u \in L$, there exists $1 \leqslant i \leqslant n$ such that $(u, u_i) \in \lambda_k$. Let us denote $I_i = \{\lg(u) \mid u \in L, (u, u_i) \in \lambda_k\}$, and $C_i = (\bigcap_{s\in \mathrm{sw}_k(u_i)} X^*sX^*) \cap \mathrm{pref}_{k-1}(u_i)X^* \cap X^*\mathrm{suff}_{k-1}(u_i) \setminus \bigcup_{t\in \mathrm{nsw}_k(u_i)} X^*tX^*$, for $i = 1, ..., n$, $n \geqslant 1$. Then $L^{\overline{A_k}} = (\bigcup_{i=1}^{n} C_i \cap \bigcup_{j\in I_i} X^j) \cup (L \cap X^{\leqslant k})$. The lower approximations $L_{\overline{A_k}}$ are of course unions of this type of sets also. The proof of statement (4) is similar to that of statement (5) of proposition 5.1.8. To prove (5), note that $\overline{A_k} \subseteq \lambda_k$ follows from the definitions, and the inclusion is proper because we can take, for every $x \in X$, $(x^k, x^{k+1}) \in \lambda_k \setminus \overline{A_k}$. The inclusion in (6) is also clear from the definition, and it is proper when $|X| > 1$, because for every $x, y \in X$, $(x^{k-1}yx^k, x^kyx^{k-1}) \in \overline{A_k} \setminus \overline{A_{k+1}}$. If $|X| = 1$ obviously $\overline{A_k} = \Delta_{X^+}$. □

The approximations under $\overline{A_k}$ are, once again, finite unions of "slices" of k-testable languages of the lengths existing in the language. Another possibility to modify A_k is to drop the length condition, as suggested at the end of [33], but in this case we would have $\overline{A_k} = \lambda_k$, and these approximations were studied in Chapter 3.

5.2.2 Modifying $M_k(X)$

Let us define $\overline{M_k}(X)$ by

$$(u,v) \in \overline{M_k}(X) \Leftrightarrow (u,v) \in M_k(X), \mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v), \mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v).$$

The relation $\overline{M_k}(X)$ thus defined is a refinement of $\lambda_k(X)$ and it is a congruence on X^+ of infinite index (every congruence class is finite). It has been studied in the context of combinatorics on words and it is known as the k-abelian equivalence. It was defined by J. Karhumäki in [16]. The statement (3) of the following lemma is mentioned without proof in that work.

5.2.2 Proposition *For every alphabet X, and $k \geqslant 1$,*

(1) $\overline{M_k} \subset \lambda_k$,

(2) $\overline{M_k} \subseteq M_k$ *and the inclusion is proper if* $|X| > 1$ *and* $k > 1$,

(3) $\varkappa = \overline{M_1} \supseteq \overline{M_2} \supseteq ...$, *and the inclusions are proper if* $|X| > 1$,

(4) $\overline{M_k} \in \mathrm{Con}(X^+) \setminus \mathrm{FCon}(X^+)$.

Proof. The inclusion in (1) follows directly from the definition of the relation, and $(x^k, x^{k+1}) \in \lambda_k \setminus \overline{M_k}$, so the inclusion is proper. The inclusion in (2) also follows from the definition, and it is proper for $|X| > 1$, $k > 1$ because $(x^{k-1}y^{k-1}x^{k-1}, y^{k-1}x^{k-1}y^{k-1}) \in M_k \setminus \overline{M_k}$. If $|X| = 1$, then $\overline{M_k} = M_k = \Delta_{X^+}$, and if $k = 1$ then $M_1 = \overline{M_1} = \varkappa$. To prove (3), suppose there exist $u, v \in X^+$, $u = x_1...x_l$, $v = y_1...y_l$, $x_i, y_i \in X$, $l > 0$, such that $(u,v) \in \overline{M_{k+1}}$ but $(u,v) \notin \overline{M_k}$. As $\overline{M_{k+1}} \subseteq \overline{A_{k+1}} \subseteq \overline{A_k}$, we know that $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$, $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$ and $\mathrm{sw}_k(u) = \mathrm{sw}_k(v)$. Then, there must be a word $w \in \mathrm{sw}_k(u)$ such that $\mu(w,u) = n$, $\mu(w,v) = m$ and $n > m$ (the case $m > n$ is similar). Firstly, we suppose that $w \neq \mathrm{suff}_k(u)$ and $w \neq \mathrm{pref}_k(u)$. Let us order the set of the n occurrences of w in u, denote by i' the place corresponding to the i^{th} appearance of w in u, and by $w_{i'}$ the appearance of w in the place i', that is to say $u = x_1 ... x_{i'-1} w_{i'} x_{i'+k} ... x_l$. For example, if $u = bbabab$ and $w = ba$, then $1' = 2$, $2' = 4$ and $u = bw_2w_4b$.

As $(u, v) \in \overline{M_{k+1}}$, the words $w_{1'}x_{1'+k}, ..., w_{m'}x_{m'+k}$ must occur at m different places $\{j_1, ..., j_m\} \subseteq \{2, ..., l-k\}$ in v. Now, the word $w_{(m+1)'}x_{(m+1)'+k}$ must also be one of the words $w_{1'}x_{1'+k}, ..., w_{m'}x_{m'+k}$ and occurs in a place $j \in \{j_1, ..., j_m\}$ at v, because $\overline{M_{k+1}} \subseteq \overline{A_{k+1}}$, but this would imply that $\mu(wx_{(m+1)'+k}, u) > \mu(wx_{(m+1)'+k}, v)$ that is a contradiction. Therefore, $n > m$ cannot hold. Similarly, $n < m$ is not possible, and thus $n = m$.

Now, if $w = \text{suff}_k(u)$ (and $w = \text{pref}_k(u)$) we can apply the same argument to the remaining $n-1 > m-1$ $(n-2 > m-2)$ occurrences of w in u. The fact that $\overline{M_1} = \varkappa$ is obvious. As for every $x, y \in X$, $(y^{k+1}xy^k, y^{k-1}xy^{k+2}) \in \overline{M_k} \setminus \overline{M_{k+1}}$, the inclusions are strict for $|X| > 1$. If $|X| = 1$, then again $\overline{M_k} = \Delta_{X^+}$. Statement (4) follows directly from the definition of $\overline{M_k}$. □

To describe the classes defined by the relation $\overline{M_k}$, we start by observing that $[u]_{\overline{M_k}} \subseteq c(u)$ for every $u \in X^+$, because $\overline{M_k} \subseteq \varkappa$. That is to say, the elements of $[u]_{\overline{M_k}}$ are permutations of u. For any word $u = x_1...x_n \in X^+$, and any permutation $\sigma \in S_n$, let $\sigma(u) = x_{\sigma^{-1}(1)}...x_{\sigma^{-1}(n)}$. For example, if $u = abaab = x_1...x_5$ and

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 4 & 5 \end{pmatrix},$$

then $\sigma(u) = x_{\sigma^{-1}(1)}...x_{\sigma^{-1}(5)} = x_3x_1x_2x_4x_5 = aabab$.

To ensure that $\sigma(u)$ has the same $(k-1)$-prefix ($(k-1)$-suffix) as u, the permutation σ must satisfy $x_{\sigma^{-1}(1)} = x_1, x_{\sigma^{-1}(2)} = x_2, ..., x_{\sigma^{-1}(k-1)} = x_{k-1}$ $(x_{\sigma^{-1}(n)} = x_n, ..., x_{\sigma^{-1}(n-k+2)} = x_{n-k+2})$. Similarly, for $\sigma(u)$ to have the same k-subwords as u, $x_{\sigma^{-1}(j+1)}x_{\sigma^{-1}(j+2)}...x_{\sigma^{-1}(j+k-1)} = x_{\sigma^{-1}(j)+1}x_{\sigma^{-1}(j)+2}...x_{\sigma^{-1}(j)+k-1}$ must hold for every $1 \leqslant j \leqslant n-k+1$.

For example, if $k = 2$, the $\sigma(u)$ given above satisfies the conditions for $u = abaab$: $x_{\sigma^{-1}(1)} = x_3 = a = x_1, x_{\sigma^{-1}(5)} = x_5 = b$ and $x_{\sigma^{-1}(1)+1} = x_4 = x_1 = x_{\sigma^{-1}(2)}, x_{\sigma^{-1}(2)+1} =$

$x_2 = x_{\sigma^{-1}(3)}, x_{\sigma^{-1}(3)+1} = x_3 = x_4 = x_{\sigma^{-1}(4)}, x_{\sigma^{-1}(4)+1} = x_4 = x_1 = x_{\sigma^{-1}(5)}$.

Let us define $T_k(u) \subseteq S_n$ as the set of permutations $\sigma \in S_n$ such that

- $x_{\sigma^{-1}(1)} = x_1, ..., x_{\sigma^{-1}(k-1)} = x_{k-1}$,
- $x_{\sigma^{-1}(n)} = x_n, ..., x_{\sigma^{-1}(n-k+2)} = x_{n-k+2}$, and
- $x_{\sigma^{-1}(j)+1} = x_{\sigma^{-1}(j+1)}, ..., x_{\sigma^{-1}(j)+k-1} = x_{\sigma^{-1}(j+k-1)}$, for all $1 \leqslant j \leqslant n-k+1$.

5.2.3 Lemma *For any word $u \in X^+$, $[u]_{\overline{M_k}} = \{u\}$ if $n \leqslant 2k-1$, and $[u]_{\overline{M_k}} = \{\sigma(u) \mid \sigma \in T_k(u)\}$ for $n > 2k-1$.*

Proof. The first statement was proved in [25]. Let us suppose that $u = x_1...x_n$ and $v = y_1...y_n \in [u]_{\overline{M_k}}$. To see that $v = \sigma(u)$ for some $\sigma \in T_k(u)$, let us define the permutation $\sigma \in S_n$ by $\sigma(i) = min\{j \mid y_j...y_{j+k-1} = x_i...x_{i+k-1}, \forall l < i, j \neq \sigma(l)\}$ for $i \leqslant n-k+1$ and $\sigma(i) = i$ for $n-k+1 < i \leqslant n$. It is clear that σ thus defined belongs to $T_k(u)$. The converse is true by the definition of $T_k(u)$. □

5.2.4 Proposition *For any $k \geqslant 1$, the family $\{\overline{M_k}(X)\}_X$ is consistent.*

Proof. Let us consider a morphism $\varphi : X^+ \to Y^+$, and two $\overline{M_k}$-related words $u, v \in X^+$. The facts that $\overline{M_k} \subseteq \varkappa$ and $\lg(u) = \lg(v)$ imply $\lg(u\varphi) = \lg(v\varphi)$. The conditions $\mathrm{pref}_{k-1}(u\varphi) = \mathrm{pref}_{k-1}(v\varphi)$ and $\mathrm{suff}_{k-1}(u\varphi) = \mathrm{suff}_{k-1}(v\varphi)$ are obvious. If $w \in \mathrm{sw}_k(u\varphi)$, then there must be a word $u' \in \mathrm{sw}_j(u)$, $j \leqslant k$, such that $w \in \mathrm{sw}_k(u'\varphi)$. Now, $\mathrm{sw}_j(u) = \mathrm{sw}_j(v)$ implies $u' \in \mathrm{sw}_j(v)$ and hence $w \in \mathrm{sw}_k(u'\varphi) \subseteq \mathrm{sw}_k(v\varphi)$. The proof of the converse inclusion $\mathrm{sw}_k(v\varphi) \subseteq \mathrm{sw}_k(u\varphi)$ is similar. If $w \in X^k$, $\mu(w, u\varphi) = n$, and we order the occurrences of w from the left, then for each $i \in [1, ..., n]$ there are words $s_i, u_i, t_i \in X^*$ such that $u = s_i u_i t_i$, $\lg(s_1) < \lg(s_2) < ... < \lg(s_n)$ and the i^{th} occurrence of w in $u\varphi$ appears in $u_i\varphi$, $u_i \in X^{\leqslant k}$. Since $\mu(u_i, u) = \mu(u_i, v)$ for every $i \in [1, ..., n]$, u_i appears in v at least as many times as it appears in the sequence $u_1, u_2, ..., u_n$, and hence $v\varphi$ contains at least n occurrences of w. Hence, $\mu(w, u\varphi) \leqslant \mu(w, v\varphi)$. By interchanging the roles of u and v, the equality $\mu(w, u\varphi) = \mu(w, v\varphi)$ is obtained. We can then conclude that $\overline{M_k}(X) \subseteq \varphi \circ \overline{M_k}(Y) \circ \varphi^{-1}$ □

The results in chapter 3 (section 3.5), apply to $\{\overline{M_k}(X)\}_X$ as the family $\{\theta \in \mathrm{FCon}(X^+) \mid \overline{M_k}(X) \subseteq \theta\}_X$ is a pseudo-principal +-filter and the corresponding +-variety $\overline{\mathcal{M}_k}$ is a pseudo-principal +-variety. Moreover, as $\overline{M_k}(X)$ has regular (finite) classes, according to Corollary 3.5.14, the following statement holds true.

5.2.5 Proposition *For every $k \geqslant 1$, a regular language $L \subseteq X^+$ has a least upper $\overline{\mathcal{M}_k}$-approximation if and only if $L^{\overline{M_k}(X)}$ is a regular language.*

5.2.3 Modifying $J_k(X)$

In [33] the authors propose to extend the above definitions to scattered subwords. For this purpose, let us define the relation $\overline{J_k}(X)$ on X^+ by

$$(u,v) \in \overline{J_k}(X) \Leftrightarrow (u,v) \in J_k(X), \text{ and } \lg(u) = \lg(v).$$

The relation $\overline{J_k}(X)$ is a refinement of $J_k(X)$, the congruence associated to piecewise k-testable languages introduced in Chapter 3, and it is a congruence of infinite index on X^+. The family $\{\overline{J_k}(X)\}_X$ is not consistent, because it does not satisfy $\overline{J_k}(X) \subseteq \varkappa$. For example when $|X| > 1$, $(aabab, abbab) \in \overline{J_2}(X) \setminus \varkappa$. We summarize some of its properties in the following

5.2.6 Proposition *For any alphabet X, $u \in X^+$, $L \subseteq X^+$, and $k \geqslant 1$,*

(1) $\overline{J_k} \subset J_k$,

(2) $\overline{J_{k+1}} \subseteq \overline{J_k}$, *and the inclusion is proper if* $|X| > 1$,

(3) $[u]_{\overline{J_k}} = \{u\}$ *if* $\lg(u) \leqslant 2k-1$,

$[u]_{\overline{J_k}} = (\bigcap_{s \in \mathrm{ssw}_k(u)} s \circ X^* \setminus \bigcup_{t \in X^k \setminus \mathrm{ssw}_k(u)} t \circ X^*) \cap X^{\lg(u)}$ *if* $\lg(u) > 2k-1$,

(4) *the rough approximations of a language* L *under* $\overline{J_k}$ *are finite unions of sets of the form* $C \cap \bigcup_{j\in I} X^j$ *where* C *is a piecewise* k*-testable language and* $I \subseteq \mathbb{N}$*, and*

(5) *if* L *is context-free, then the sets* I *in (4) form arithmetic sequences and the rough approximations of* L *under* $\overline{J_k}$ *are regular.*

Proof. The inclusion in (1) follows from the definition, and it is proper because $(x^{k+1}, x^k) \in J_k \setminus \overline{J_k}$ for every $x \in X$. To prove (2), observe that $(u, v) \in \overline{J_{k+1}} \subseteq J_{k+1} \subseteq J_k$, implies $(u, v) \in J_k$, and as $\lg(u) = \lg(v)$, we can conclude that $(u, v) \in \overline{J_k}$. When $|X| = 1$, $\overline{J_k} = \Delta_{X^+}$, and if $|X| > 1$, the inclusion is proper because for every $x, y \in X$, $(y^{k+1}x^k, y^k x^{k+1}) \in \overline{J_k} \setminus \overline{J_{k+1}}$. The fact that a word of length less or equal than $2k - 1$ is determined by its set of scattered subwords of length k, was proved in [42] (cf. also [25]), thus $[u]_{\overline{J_k}} = \{u\}$ if $\lg(u) \leqslant 2k - 1$. The characterization of the classes of a word of length greater than $2k - 1$ follows from the definition of the relation. The fact (4) is a consequence of (3). The proof of (5) is the same, mutatis mutandis, as the one of statement (5) in proposition 5.1.8. □

5.3 Some comments on accuracy

The length condition imposed in the definitions of the relations in this chapter tends to make the approximations better when the original language has gaps in its density function, for example the ρ_2-approximation of the language $R = (X^2)^+$ is $R^{\rho_2} = X^2X^*$ but the P_2-approximation gives immediately $R^{P_2} = R$. The ρ_2-accuracy does not exist in this case (cf. example 4.2.6). However, when the language has no gaps, the length condition makes little difference in the approximations. We show some examples of these situations and how our measures of accuracy can reflect them.

5.3.1 Example Let $X = \{0, 1\}$.

(1) The language $R_1 = \{w \in X^+ \mid |w|_0 \equiv_2 0\}$ has density $d_{R_1}(m) = 2^{m-1}$. The approximations $R_1^{\rho_3} = \{1, 00, 11\} \cup X^3X^*$, $R_1^{P_3} = \{1, 00, 11, 001, 010, 100, 111\} \cup$

X^3X^+, have densities $d_{R_1^{\rho_3}}(m) = d_{R_1^{P_3}}(m) = 2^m$ for every $m \geqslant 4$, and the accuracies are $Acc(R_1, P_3) = Acc(R_1, \rho_3) = \lim_{m\to\infty} \frac{2^{m-1}}{2^m} = 1/2$. The equality in the accuracy, reflects the fact that both approximations are similar in this no-gaps language.

(2) An interesting case showing how different the behavior of accuracies of approximations under ρ_k and P_k can be, is illustrated by $R = \{00\} \cup 11^+$. We have $R^{P_2} = \{00\} \cup 11X^*$ and $R^{\rho_2} = 00X^* \cup 11X^*$, thus $d_{R^{P_2}}(m) = 2^{m-2} < 2 \cdot 2^{m-2} = d_{R^{\rho_2}}(m)$ for every $m > 2$. It is clear that both $Acc(R, \rho_2)$ and $Acc(R, P_2)$ are zero, but the fact that the P_2-approximation is better is shown by $\lim_{m\to\infty} |R^{P_2} \cap X^m|/|R^{\rho_2} \cap X^m| = 1/2$. This kind of phenomenon occurs when there is a word u of length k in the language that does not appear as a k-prefix in any other word of the language. Then, uX^* is added to R^{ρ_k} but just $\{u\}$ is added to R^{P_k}.

(3) This example is a variation of the previous one, in which the accuracy, as we define it, does show the difference between ρ_k and P_k-approximations. Let $R = \{00\} \cup 11R_1$, where R_1 is the language in the item (1) above. In this case we have again $R^{P_2} = \{00\} \cup 11X^*$ and $R^{\rho_2} = 00X^* \cup 11X^*$, but now for every $m \geqslant 3$, $d_R(m) = 2^{m-3}$, thus $Acc(R, \rho_2) = \lim_{m\to\infty} \frac{2^{m-3}}{2\cdot 2^{m-2}} = 1/4$ and $Acc(R, P_2) = \lim_{m\to\infty} \frac{2^{m-3}}{2^{m-2}} = 1/2$.

(4) Consider now a language with gaps. The language $R = (111 + 000)^+$, is better approximated by P_3 than by ρ_3. Indeed, $R^{P_3} = (111 + 000)(X^3)^*$, and $R^{\rho_3} = (111+000)X^*$. The accuracy in the limit is in both cases 0, but for every $n \geqslant 2$, $Acc(R, P_3, 3n) = \frac{(2^{n+1}-2)7}{2^{3n+1}-2} > \frac{2^{n+1}-2}{2^{3n-1}-2} = Acc(R, \rho_3, 3n)$.

(5) Finally, we give an example comparing the accuracy attained in λ_k and in $\overline{A_k}$-approximations. Let $R = (00)^+(11)^+$. We obtain $R^{\lambda_2} = 00^+11^+$ and $R^{\overline{A_2}} = 00^+11^+ \cap X^2(X^2)^+$. To calculate the accuracies we consider the even and the odd lengths separately because neither the language, nor the $\overline{A_2}$-approximation,

have words of odd length. For every $n \geqslant 2$,

$$\lim_{n\to\infty} Acc(R, \lambda_2, 2n) = \lim_{n\to\infty} \frac{(1+2+...+n-1)}{(1+2+...+2n-3)} = \lim_{n\to\infty} \frac{n(n-1)}{(2n-3)(2n-2)} = 1/4,$$

$$\lim_{n\to\infty} Acc(R, \lambda_2, 2n+1) = \lim_{n\to\infty} \frac{(1+2+...+n-1)}{(1+2+...+2n-2)} = \lim_{n\to\infty} \frac{n(n-1)}{(2n-2)(2n-1)} = 1/4$$

and

$$\lim_{n\to\infty} Acc(R, \overline{A_2}, 2n) = \lim_{n\to\infty} \frac{(1+2+...+n-1)}{(1+3+...+2n-3)} = \lim_{n\to\infty} \frac{n(n-1)}{(n-1)(2n-2)} = 1/2,$$

$$\lim_{n\to\infty} Acc(R, \overline{A_2}, 2n+1) = \lim_{n\to\infty} \frac{(1+2+...+n-1)}{(1+3+...+2n-3)} = \lim_{n\to\infty} \frac{n(n-1)}{(n-1)(2n-2)} = 1/2,$$

Therefore, the accuracies $Acc(R, \lambda_2)$ and $Acc(R, \overline{A_2})$ exist, and the inequality $Acc(R, \lambda_2) = 1/4 < 1/2 = Acc(R, \overline{A_2})$ shows the fact that a language with gaps like this, may be better approximated by $\overline{A_2}$ than by λ_2.

In general, we can state that

5.3.2 Proposition *Let $R \subseteq X^+$ and $k \geqslant 1$. Whenever the involved limits exist,*

(1) $Acc(R, \rho_k) \leqslant Acc(R, P_k)$,

(2) $Acc(R, J_k) \leqslant Acc(R, \overline{J_k}))$, *and*

(3) $Acc(R, \lambda_k) \leqslant Acc(R, \overline{A_k}) \leqslant Acc(R, \overline{M_k})$.

Proof. To prove (1) it is enough to observe that $P_k \subseteq \rho_k$ implies $R^{P_k} \cap X^{\leqslant m} \subseteq R^{\rho_k} \cap X^{\leqslant m}$, and hence $Acc(R, P_k, m) \geqslant Acc(R, \rho_k, m)$ for every $m \geqslant 1$, therefore $Acc(R, P_k) \geqslant Acc(R, \rho_k)$. The proof of the rest of the inequalities is completely similar. □

5.4 Concluding remarks

As the work of Păun, Polkowski and Skowron [33] seems to be one of the few predecessors of our work on rough approximation of languages in the literature, in this chapter we studied their approach in some detail. The indiscernibility relations they define are not always congruences, in some cases they are not even equivalence relations, and they have all infinite index. We performed the modifications needed to obtain congruences, described the general features of these relations, and in most cases we gave characterizations of the equivalence classes under them. Some are related to the families of k-definite, reverse k-definite, k-testable and piecewise k-testable studied previously, so that we compared the rough approximations they generate with those of the mentioned families. The characterization of the classes of M_k is a combinatorial problem that remains to be solved. It is worth noting the case of $\{\overline{M}_k\}_X$, that turns out to be a consistent family, giving a new pseudo-principal +-variety to which we applied our previous results. Finally we showed some examples of the accuracy of the rough approximations given by the new relations comparing them with the old ones. In general, we observed that the approximations under relations with the length condition tend to be better when the original language has gaps in its density. A full study of approximations under tolerance relations remains to be done, but this task would certainly require a different theoretical framework.

Bibliography

[1] J. Almeida, *Finite Semigroups and Universal Algebra*, World Scientific, Singapore, 1995.

[2] D. Angluin and C. H. Smith, Inductive inference:theory and methods, *ACM Computing Surveys* **15**, No.3 (1983), 237-270.

[3] B. Badban, M. Torabi Dashti, Semi-linear Parikh Images of Regular Expressions via Reduction, *Lecture Notes in Computer Science* **6281**, (2010), 653-664.

[4] J. Berstel, Sur la densité asymptotique de langages formels, *Automata, Languages, and Programming*, M. Nivat (ed.), North Holland, 1973, 345-358.

[5] P. E. Black (ed.), *Dictionary of Algorithms and Data Structures* [online], U.S. National Institute of Standards and Technology, 2005.

[6] J. A. Brzozowski, Canonical regular expressions and minimal state graphs of definite events, *Proc. of the Symposium on Mathematical Theory of Automata* **12** (1963), Brooklyn, New York, 529-561.

[7] S. Burris and H. P. Sankappanavar, *A Course in Universal Algebra*, Springer-Verlag, New York, 1981.

[8] B. Cordy and K. Salomaa, On the existence of regular approximations, *Theoretical Computer Science* **387** (2007), 125-135.

[9] N. Chomsky and M. P. Schützenberger, The algebraic theory of context-free languages. *Computer Programming and Formal Languages*, P. Braffort and D. Hirschberg (eds.), North Holland, 1963, 118-161.

[10] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, 2nd. edition, Cambridge University Press, Cambridge, 2002.

[11] S. Eilenberg, *Automata, Languages, and Machines*, Vol. A, Academic Press, New York, 1974.

[12] S. Eilenberg, *Automata, Languages, and Machines*, Vol. B, Academic Press, New York, 1976.

[13] S. Ginsburg and E.H. Spanier, Bounded regular sets, *Proc. American Mathematical Soc.* **17** (1966), 1043-1049.

[14] A. Ginzburg, About some properties of definite, reverse-definite and related automata, *IEEE Trans. Electronic Computers* **EC-15** (1966), 809-810.

[15] R.C. Gonzalez and M. G. Thomason, *Syntactic Pattern Recognition: An Introduction*, Addison-Wesley, Reading, MA, 1978.

[16] J. Karhumäki, Generalized Parikh Mappings and Homomorphisms, *Information and Control*, **47 (3)** (1980), 155-165.

[17] S. C. Kleene, *Representation of events in nerve nets and finite automata*, in C. E. Shannon and J. McCarthy (eds.) Automata Studies, Princeton University Press, Princeton, NJ, 1956, 3-42.

[18] D. Knuth, Big Omicron and big Omega and big Theta, *ACM SIGACT News*, **8(2)** (1976), 18-24.

[19] A. Kornai, Quantitative comparison of languages, *Grammars* **1(2)** (1988), 155-165.

[20] C. Koutschan, Regular languages and their generating functions: the inverse problem, *Theoretical Computer Science*, **391** (2008), 65-74.

[21] J. Kozik, Conditional densities of regular languages, *Electronic Notes in Theoretical ComputerScience* **140** (2005), 67-79.

[22] D. C. Kozen, *Automata and Computability*, Springer, New York, 1997.

[23] M. V. Lawson, *Finite Automata*, Chapman & Hall/CRC, Boca Raton, 2004.

[24] M. Lothaire, *Combinatorics on words*, Addison-Wesley, 1983.

[25] J. Maňuch, Characterization of a word by its subwords, *Developments in Language Theory*, G. Rozenberg and W.Thomas (eds.), World Scientific, 2000, 210-219.

[26] S. Marcus, On the length of words. *Jewels are Forever. Contributions on Theoretical Computer Science in Honour of Arto Salomaa*, Juhani Karhumäki, Hermann Maurer, Gheorghe Paun (eds.), Springer, London, 1999, 194-203.

[27] G. Martín and M. Steinby, Rough Approximations in varieties of regular languages, *Fundamenta Informaticae* **112(4)** (2011), 281-303.

[28] G. Martín Torres, On the Accuracy of Rough Approximations of Regular Languages, *Fundamenta Informaticae* **132** (2014), 1-13.

[29] R. McNaughton and S. Papert, *Counter-free automata*, MIT Press, Cambridge, Massachusetts, 1971.

[30] L. Miclet, *Structural Methods in Pattern Recognition*, Springer-Verlag, New York, 1986.

[31] E. Orłowska (ed.), *Incomplete Information: Rough Set Analysis*, Studies in Fuzziness and Soft Computing, Physica-Verlag, Heidelberg, 1998.

[32] Gh. Păun, L. Polkowski and A. Skowron, Rough-set-like approximations of context-free and regular languages, *Proc. Information Processing and Management of Uncertainity on Knowledge Based Systems, (Proc. IPMU-96)* **2** (1996), 891-895.

[33] Gh. Păun, L. Polkowski and A. Skowron, Rough set of approximations of languages, *Fundamenta Informaticae* **32** (1997), 149-162.

[34] Z. Pawlak, Rough sets, *International Journal of Computer and Information Science* **11** (1982), 341-356.

[35] Z. Pawlak, Rough sets, *Theoretical aspects of reasoning about data*, Kluwer Academic Publishers, Dordrecht, 1991.

[36] M. Perles, M. O. Rabin and E. Shamir, The theory of definite automata, *IEEE Trans. Electronic Computers* **EC-12** (1963), 233-243.

[37] V. Piirainen, Principal varieties of finite congruences, *TUCS Technical Report* No. 874, Turku Centre for Computer Science, Turku, 2008.

[38] J.-E. Pin, *Varieties of formal languages*, North Oxford Academic Publ., London, 1986.

[39] J.-E. Pin, Syntactic semigroups, *Handbook of Formal Languages, Vol.1*, G. Rozenberg and A. Salomaa, (eds.), Springer, Berlin 1997, 679-746.

[40] L. Polkowski, *Rough sets. Mathematical Foundations*, Advances in Soft Computing, Physica-Verlag, Heidelberg, 2002.

[41] H. Seidl, T. Schwentick , A. Muscholl, and P. Habermeh, Counting in Trees for Free, *Lecture Notes in Computer Science* **3142**, (2004), 1136-1149

[42] I.Simon, Piecewise testable events, *Proceedings of the 2nd GI Conference on Automata Theory and Formal Languages, Lecture Notes in Computer Sciences* **33** (1975), 214-222.

[43] A. M. Shur, Combinatorial Complexity of Rational Languages, *Diskretnyj Analiz i Issledovanie Operatsij.* **12:2** (2005) 78-99.

[44] M. Spivak, *Calculus*, University Press, Cambridge, 2006.

[45] M. Steinby, A theory of tree language varieties, in: *Tree Automata and Languages* M. Nivat and A. Podelski (eds.), North-Holland, Amsterdam, 1992, 57-81.

[46] M. Steinby, Algebraic classifications of regular tree languages, in: *Structural Theory of Automata, Semigroups, and Universal Algebra* Kudryavtsev and Rosenberg (eds.), Springer, Dordrecht 2005, 381-432.

[47] A. Szilard, S. Yu, K. Zhang and J. Shallit, Characterizing Regular Languages with Polynomial Densities, *Lecture Notes in Computer Science, Mathematical Foundations of Computer Science* **629**, Springer-Verlag, London, 1992, 494-503.

[48] D. Thérien, Classification of regular congruences, Ph.D. Thesis, *Research Report CS-80-19*, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1980.

[49] D. Thérien, Classification of finite monoids: the language approach, *Theoretical Computer Science* **14** (1981), 195–208.

[50] A. To, Parikh Images of Regular Languages: Complexity and Applications, in: *CoRR, abs/1002.1464*, (2010).

[51] S. Yu, Regular languages, in: *Handbook of Formal Languages*, Vol. 1 G. Rozenberg and A. Salomaa (eds.), Springer-Verlag, Berlin, 1997, 41-110.