Resynchronizing Classes of Word Relations

María Emilia Descotte, Diego Figueira, Gabriele Puppis

Resynchronizing Classes of Word Relations*

María Emilia Descotte[2], Diego Figueira[1,2], and Gabriele Puppis[1,2]

1 **CNRS**
2 **LaBRI, Université de Bordeaux**

Abstract

A natural approach to defining binary word relations over a finite alphabet $\mathbb{A}$ is through two-tape finite state automata, whose runs are described by regular languages L over $\{1,2\} \times \mathbb{A}$, where (i,a) is interpreted as reading letter a from tape i. Accordingly, a word $w \in L$ denotes the pair $(u_1,u_2) \in \mathbb{A}^* \times \mathbb{A}^*$ in which u_i is the projection of w onto i-labelled letters. While this formalism defines the well-studied class of Rational relations (*a.k.a.* non-deterministic finite state transducers), enforcing restrictions on the reading regime from the tapes, which we call *synchronization*, yields various sub-classes of relations. Such synchronization restrictions are imposed through regular properties on the projection of the language onto $\{1,2\}$. In this way, for each regular language $C \subseteq \{1,2\}^*$, one obtains a class $\textsc{Rel}(C)$ of relations. Regular, Recognizable, and length-preserving relations are all examples of classes that can be defined in this way.

We study the problem of containment for synchronized classes of relations: given $C, D \subseteq \{1,2\}^*$, is $\textsc{Rel}(C) \subseteq \textsc{Rel}(D)$? We show a characterization in terms of C and D which gives a decidability procedure to test for class inclusion. This also yields a procedure to re-synchronize languages from $\{1,2\} \times \mathbb{A}$ preserving the denoted relation whenever the inclusion holds.

1 Introduction

Our objects of study are relations of finite words, that is, binary relations $R \subseteq \mathbb{A}^* \times \mathbb{A}^*$ for a finite alphabet $\mathbb{A}$. The study of relations dates back to the works of Büchi, Elgot, Mezei, and Nivat in the 1960s [3, 7, 12], with much subsequent work done later (*e.g.*, [1, 5]). Most of the investigations focused on extending the standard notion of regularity from languages to relations. This effort has followed the long-standing tradition of using equational, operational, and descriptive formalisms – that is, finite monoids, automata, and regular expressions – for describing relations, and gave rise to three different classes of relations: the *Recognizable*, the *Automatic* (*a.k.a. Regular* [1] or *Synchronous* [5]), and the *Rational* relations.

The above classes of relations can be seen as three particular examples of a much larger (in fact infinite) range of possibilities, where relations are described by special languages over extended alphabets, called *synchronizing languages* [9]. Intuitively, the idea is to describe a binary relation by means of a two-tape automaton with two heads, one for each tape, which can move independently one of the other. In the basic framework of synchronized relations, one lets each head of the automaton to either move right or stay in the same position. In addition, one can constrain the possible sequences of head motions by a suitable regular language $C \subseteq \{1,2\}^*$. In this way, each regular language $C \subseteq \{1,2\}^*$ induces a class of binary relations, denoted $\textsc{Rel}(C)$. For example, the class of Recognizable, Automatic, and Rational relations are captured, respectively, by the languages $C_{\text{Rec}} = \{1\}^* \cdot \{2\}^*$, $C_{\text{Aut}} = \{12\}^* \cdot \{1\}^* \cup \{12\}^* \cdot \{2\}^*$, and $C_{\text{Rat}} = \{1,2\}^*$. In general, the correspondence between a language $C \subseteq \{1,2\}^*$ and the induced class $\textsc{Rel}(C)$ of synchronized relations is not one-to-one: it may happen that different languages C, D induce the same class of

* Work supported by ANR project DELTA, grant ANR-16-CE40-0007, and LIA INFINIS.

synchronized relations. There are thus fundamental questions that arise naturally in this framework: *When two classes of synchronized relations coincide, and when is one contained in the other?* Our contribution is a precise algorithmic answer to these types of questions.

More concretely, given a binary alphabet $\mathbb{2} = \{1, 2\}$ and another finite alphabet $\mathbb{A}$, a word $w \in (\mathbb{2} \times \mathbb{A})^*$ is said to *synchronize* the pair $(w_1, w_2) \in \mathbb{A}^* \times \mathbb{A}^*$ if, for both $i = 1, 2$, w_i is the projection of w on $\mathbb{A}$ restricted to the positions marked with i. For short, we denote this by $[\![w]\!] = (w_1, w_2)$ —*e.g.*, $[\![(1, a)(1, b)(2, b)(1, a)(2, c)]\!] = (aba, bc)$. According to this definition, every word over $\mathbb{2} \times \mathbb{A}$ synchronizes a pair of words over $\mathbb{A}$, and every pair of words over $\mathbb{A}$ is synchronized by (perhaps many) words over of $\mathbb{2} \times \mathbb{A}$. This notion is readily lifted to languages: a language $L \subseteq (\mathbb{2} \times \mathbb{A})^*$ synchronizes the set of pairs (*i.e.*, the relation) $[\![L]\!] = \{[\![w]\!] \mid w \in L\} \subseteq \mathbb{A}^* \times \mathbb{A}^*$. For example, $[\![((1, a)(2, a) \cup (1, b)(2, b))^*]\!]$ denotes the equality relation over $\mathbb{A} = \{a, b\}$.

In this setup, one can define classes of relations by restricting the set of admitted synchronizations. The natural way of doing so is to fix a language $C \subseteq \mathbb{2}^*$, called *control language*, and let L vary over all regular languages over the extended alphabet $\mathbb{2} \times \mathbb{A}$, so that the projection of L onto $\mathbb{2}$ is contained in C. Thus, for every regular $C \subseteq \mathbb{2}^*$, there is an associated class $\textsc{Rel}(C)$ of C-controlled relations, namely, relations synchronized by regular languages $L \subseteq (\mathbb{2} \times \mathbb{A})^*$ whose projection onto $\mathbb{2}$ are contained in C. It is immediate to see that for all $C \subseteq D \subseteq \mathbb{2}^*$, $\textsc{Rel}(C) \subseteq \textsc{Rel}(D)$. However, this is not a necessary condition: while $\textsc{Rel}(C_{\text{Rec}}) = \textit{Recognizable} \subseteq \textit{Automatic} = \textsc{Rel}(C_{\text{Aut}})$, we have $C_{\text{Rec}} \not\subseteq C_{\text{Aut}}$. Moreover, as we have mentioned earlier, different control languages may induce the same class of synchronized relations. For example, once again, the class of Recognizable relations is induced by the control language $C_{\text{Rec}} = \{1\}^*\{2\}^*$, but also by $C'_{\text{Rec}} = \{1\}^*\{2\}^*\{1\}^*$, and the class of Automatic relations is induced by $C_{\text{Aut}} = \{12\}^* \cdot \{1\}^* \cup \{12\}^* \cdot \{2\}^*$, or equally by $C'_{\text{Aut}} = \{21\}^* \cdot \{1\}^* \cdot \{2\}^*$.

This 'mismatch' between control languages and induced classes of relations gives rise to the following algorithmic problem:

Class Containment Problem	
Input:	Two regular languages $C, D \subseteq \mathbb{2}^*$
Output:	Is $\textsc{Rel}(C) \subseteq \textsc{Rel}(D)$?

Note that the above problem is different from the (C, D)-membership problem on synchronized relations, which consists in deciding whether $R \in \textsc{Rel}(D)$ for any given $R \in \textsc{Rel}(C)$, and which can be decidable or undecidable depending on C, D [4]. The Class Containment Problem can be seen as the problem of whether *every* C-controlled regular language L has a D-controlled regular language L' so that $[\![L]\!] = [\![L']\!]$. Our main contribution is a procedure for deciding this problem:

▶ **Main Theorem.** *The Class Containment Problem is decidable.*

In addition, our results show that, for positive instances (C, D), one can effectively transform any regular C-controlled language L into a regular D-controlled language L' so that $[\![L]\!] = [\![L']\!]$. By 'effectively transform' we mean that one can receive as input an automaton (or a regular expression) for L and produce an automaton (or a regular expression) for L'.

Related work. The formalization of a framework in which one can describe classes of word relations by means of synchronization languages is quite recent [9]. The class containment problem was only addressed for the classes of Automatic, length-preserving, and Rational relations, for which several characterizations have been proposed [9]. The formalism of synchronizations has been extended beyond rational relations by means of semilinear constraints [8] in the context of path querying languages for graph databases.

The paper [2] studies relations with origin information, as induced by non-deterministic (one-way) finite state transducers. Origin information can be seen as a way to describe a synchronization between input and output words – somehow in the same spirit of our synchronization languages – and was exploited to recover decidability of the equivalence problem for transducers. The paper [10] pursues further this principle by studying "distortions" of the origin information, called resynchronizations. Despite the similar terminology and the connection between origins and synchronizing languages, the problems studied in [2, 10] are of rather different nature than our Class Containment Problem.

Organization. After the preliminaries on regular languages and subclasses of them that are relevant for our characterization, we define in Section 3 the framework of synchronized relations, and exhibit the fundamental properties. Section 4 provides a roadmap for the key ingredients of our characterization, namely: (i) a decomposition of regular control languages into so-called '*simple*' control languages, (ii) a characterization of containment in the special case of simple languages, and (iii) a generalization of the characterization to unions of simple languages. Sections 5, 6 and 7 contain the technical details for the previous main ingredients. Finally, Section 8 includes a discussion on the computability and complexity consequences that stem from the characterization. The omitted details of the proofs are in the Appendix.

2 Preliminaries

We denote by $\mathbb{N}, \mathbb{Q}$ the sets of non-negative integers and rationals. We use standard interval notation as in, for example, $(a, b]_{\mathbb{Q}} = \{c \in \mathbb{Q} \mid a < c \leq b\}$. $\mathbb{A}, \mathbb{B}$ denote arbitrary finite alphabets, and $\mathbb{2}$ the a special binary alphabet $\{1, 2\}$.

Words and shuffles. For a word $w \in \mathbb{A}^*$, $|w|$ is its length, and $|w|_a$ is the number of occurrences of symbol a in w. We denote by $w[i, j]$ the factor of w between positions i and j (included), for $1 \leq i \leq j \leq |w|$, and we write $w[i]$ for $w[i, i]$. We will also make use of the **shuffle** operation, which maps a finite set of words $w_1, \ldots, w_n$ to the language $\mathit{shuffle}\{w_1, \ldots, w_n\}$ of all words w for which there is a partition $I_1, \ldots, I_n$ of $[1, |w|]$ so that each w_i is the projection of w onto I_i. For example, $\mathit{shuffle}\{ab, cd\} = \{abcd, cdab, acbd, acdb, \ldots\}$.

Parikh image. The **Parikh image** of a word w over $\mathbb{A}$ is the tuple $\pi(w)$ associating each symbol $a \in \mathbb{A}$ to its number of occurrences $|w|_a$ in w. We will mostly use Parikh images for words over $\mathbb{2}^*$, which are thus pairs $\pi(w) = (|w|_1, |w|_2)$. We naturally extend this to languages by letting $\pi(L) \stackrel{\text{def}}{=} \{\pi(w) \mid w \in L\}$ $(\subseteq \mathbb{N}^2)$. For $\bar{x}, \bar{x}_1, \ldots, \bar{x}_n \in \mathbb{N}^2$, we represent the linear set $\{\bar{x} + \alpha_1\bar{x}_1 + \cdots + \alpha_n\bar{x}_n \mid \alpha_1, \ldots, \alpha_n \in \mathbb{N}\}$ in dimension 2 as a pair $\langle \bar{x}, P\rangle$, where $\bar{x} \in \mathbb{N}^2$ is the **basis** and $P = \{\bar{x}_1, \ldots, \bar{x}_n\}$ is the set of **periods**. A **semilinear** set is a finite union of linear sets.

Regular languages. We use standard notation for regular expressions without complement, namely, for expressions build up from the empty word ε and the symbols $a \in \mathbb{A}$, using the operations $\cdot$, $\cup$, and $(\,)^*$. For economy of space and clarity we also use the abbreviated notation $(\,)^k$, $(\,)^{k*}$, $(\,)^{\geq k}$, $(\,)^{<k}$, and we abuse notation by identifying regular expressions with the defined languages. For example, we may write $abbc \in a \cdot b^{\geq 2} \cdot (c \cup d)^*$, $b(ab)^* = (ba)^*b$ and $\{a, b\}^* \cdot c = (a \cup b)^* \cdot c$. For two words $u = a_1 \cdots a_n \in \mathbb{A}^*$ and $v = b_1 \cdots b_n \in \mathbb{B}^*$, we write $u \otimes v$ for the word $(a_1, b_1) \cdots (a_n, b_n) \in (\mathbb{A} \times \mathbb{B})^*$, and for two languages $U \subseteq \mathbb{A}^*$, $V \subseteq \mathbb{B}^*$, we write $U \otimes V \subseteq (\mathbb{A} \times \mathbb{B})^*$ for the set $\{u \otimes v \mid u \in U, v \in V, |u| = |v|\}$.

The **star-height** of a regular expression is the maximum number of nestings of Kleene stars $(\,)^*$. By a slight abuse of terminology, when referring to the star-height of a language, we usually mean the star-height of some regular expression that represents it (in particular,

we do not need to work with the minimum star-height over all possible expressions). Besides regular expressions, we will also work with automata, and use classical techniques on them (notably, pumping arguments). Given an accepting run γ of an automaton $\mathcal{A}$, we will often identify **cycles** in it, that is, factors that start and end in the same state, and that can thus be pumped. Such cycles are called **simple** if they do not contain proper factors that are also cycles. Moreover, to avoid mentioning explicitly an automaton for a language L and a run of it, we will call **cycle of** L (resp. **simple cycle of** L) the word spelled out by any cycle (resp. simple cycle) of any accepting run of the minimal deterministic automaton recognizing L, and denote the set of all cycles (resp. simple cycles) of L by $\mathit{cycles}(L)$ (resp. $\mathit{simple\text{-}cycles}(L)$). It is worth noting, however, that the use of a minimal deterministic automaton as a presentation of a regular language L is only to avoid ambiguity when referring to the cycles of L —in fact, our results do not depend on determinism or minimality, and can thus be applied to regular languages presented by arbitrary non-deterministic automata, without any difference in the characterizations we will present.

We say that a regular language C is **concat-star**, if it is of the form

$$C = C_1^* u_1 C_2^* u_2 \cdots C_k^* u_k, \tag{$\star$}$$

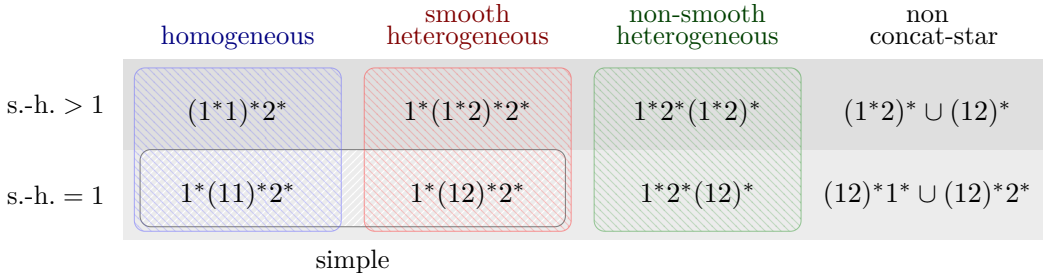
for $k \in \mathbb{N}$, words $u_1, \ldots, u_k$, and regular languages $C_1, \ldots, C_k$. Without loss of generality, we can always assume that the empty word does not belong to any of the languages C_i. The following trival decomposition lemma will be used throughout.

▶ **Lemma 1.** *Every regular language is a finite union of concat-star languages.*

The C_i^*'s from ($\star$) are called **components** of the concat-star language C. Note that (an expression of) a concat-star language as in ($\star$) has star-height 1 if and only if every C_i is finite. A component C_i^* is **1-homogeneous** if $C_i^* \subseteq 1^*$, and **2-homogeneous** if $C_i^* \subseteq 2^*$. A **homogeneous** component is an i-homogeneous component for some i. A component which is not homogeneous is called **heterogeneous** (e.g. $C_i^* = \{1,2\}^*$). It will also be convenient to distinguish a few of types of concat-star languages. We say that C is

- **heterogeneous** if it contains at least one heterogeneous component, otherwise it is **homogeneous**;
- **smooth** if every homogeneous component is a language of the form 1^{k*} or 2^{k*}, for some $k > 0$, and there are no consecutive homogeneous components;
- **simple** if it has star-height 1 and it is either homogeneous or smooth heterogeneous.

Hereafter, by "simple language" we mean simple concat-star language. The picture below summarizes the different types of control languages that we will consider, together with some separating examples.



In Section 5 we will see that the Class Containment Problem is reduced to the case of finite unions of simple languages. The latter languages thus form the basis of our characterization.

3 Synchronized relations

A **synchronization** of a pair (w_1, w_2) of words over $\mathbb{A}$ is a word over $\mathbb{2} \times \mathbb{A}$ so that the projection on $\mathbb{A}$ of positions labeled i is exactly w_i, for $i = 1, 2$ —in other words, $\mathit{shuffle}\{1^{|w_1|} \otimes w_1, 2^{|w_2|} \otimes w_2\}$ is the set of all synchronizations of (w_1, w_2). For example, the words $(1,a)(1,b)(2,a)$ and $(1,a)(2,a)(1,b)$ are two possible synchronizations of the same pair (ab, a). Every word $w \in (\mathbb{2} \times \mathbb{A})^*$ is a synchronization of a unique pair (w_1, w_2), where w_i is the sequence of $\mathbb{A}$-letters corresponding to the symbol i in the first position of $\mathbb{2} \times \mathbb{A}$. We denote such pair (w_1, w_2) by $[\![w]\!]$ and extend the notation to languages $L \subseteq (\mathbb{2} \times \mathbb{A})^*$ by $[\![L]\!] \stackrel{\text{def}}{=} \{[\![w]\!] \mid w \in L\}$.

Given a regular language $C \subseteq \mathbb{2}^*$, we define the class of ***C*-controlled relations** as

$$\textsc{Rel}(C) \stackrel{\text{def}}{=} \big\{[\![L]\!] \;\mid\; L \subseteq C \otimes \mathbb{A}^* \text{ is regular}, \mathbb{A} \text{ is some finite alphabet}\big\}.$$

A slightly different definition is possible, which restricts the class of C-controlled relations to be over a fixed alphabet $\mathbb{A}$, that is, one can define $\textsc{Rel}_{\mathbb{A}}(C) = \{[\![L]\!] \mid L \subseteq C \otimes \mathbb{A}^* \text{ regular}\}$. As far as we are concerned with comparing classes of relations controlled by different languages, the two definitions are somehow interchangeable, in the sense that containment between classes is not sensible to whether we fix or not the alphabet. For example, we will see that, for any alphabet $\mathbb{A}$ with at least two symbols, $\textsc{Rel}_{\mathbb{A}}(C) \subseteq \textsc{Rel}_{\mathbb{A}}(D)$ iff $\textsc{Rel}(C) \subseteq \textsc{Rel}(D)$.

For economy of space, we use $C \subseteq_{\textsc{Rel}} D$ and $C =_{\textsc{Rel}} D$ as shorthands for $\textsc{Rel}(C) \subseteq \textsc{Rel}(D)$ and $\textsc{Rel}(C) = \textsc{Rel}(D)$, respectively. The following properties are easy to verify.

▶ **Lemma 2.** *For every regular $C, D, C', D' \subseteq \mathbb{2}^*$,*

P1. *if $C \subseteq D$, then $C \subseteq_{\textsc{Rel}} D$;*
P2. *if $C \subseteq_{\textsc{Rel}} D$ and $C' \subseteq_{\textsc{Rel}} D'$, then $C \cdot C' \subseteq_{\textsc{Rel}} D \cdot D'$ and $C \cup C' \subseteq_{\textsc{Rel}} D \cup D'$;*
P3. *if $C \subseteq_{\textsc{Rel}} D$, then $C^* \subseteq_{\textsc{Rel}} D^*$;*
P4. *if $C \subseteq 1^*$ and $D \subseteq 2^*$, then $C \cdot D =_{\textsc{Rel}} D \cdot C$;*
P5. *if C is finite, then $C \cdot D =_{\textsc{Rel}} D \cdot C$;*
P6. *if $C \subseteq_{\textsc{Rel}} D$ then $\pi(C) \subseteq \pi(D)$; moreover, if C is finite, the converse also holds;*
P7. *if C is homogeneous concat-star, then $C =_{\textsc{Rel}} \bigcup_{i \in I} 1^{\ell_i *} 1^{k_i} 2^{\hat{\ell}_i *} 2^{\hat{k}_i}$ for a finite I;*
P8. *if C is homogeneous concat-star, $C \subseteq_{\textsc{Rel}} D$ if and only if $\pi(C) \subseteq \pi(D)$.*

Proof idea. P1 is immediate from definitions; henceforth we use it without referencing it. P2 and P3 follow readily from the following decomposition properties:

(a) For every $R \in \textsc{Rel}(C \cdot C')$, there are $R_1, \ldots, R_n \in \textsc{Rel}(C)$, $R'_1, \ldots, R'_n \in \textsc{Rel}(C')$ so that $R = \bigcup_i R_i \cdot R'_i$.
(b) For every $R \in \textsc{Rel}(C \cup C')$, thee are $R_1 \in \textsc{Rel}(C)$, $R_2 \in \textsc{Rel}(C')$ so that $R = R_1 \cup R_2$.
(c) For every $R \in \textsc{Rel}(C^*)$, there are $R_1, \ldots, R_n \in \textsc{Rel}(C)$ and $I \subseteq \{1, \ldots, n\}^*$ regular so that $R = \bigcup_{w \in I} R_{w[1]} \cdots R_{w[|w|]}$.

P4 can be verified by first decomposing any relation $R \in \textsc{Rel}(C \cdot D)$ into $\bigcup_i R_i \cdot R'_i$ as in (a), and then observing that in this case $[\![\bigcup_i R_i \cdot R'_i]\!] = [\![\bigcup_i R'_i \cdot R_i]\!]$. For P5, it is easy to see that $1 \cdot D =_{\textsc{Rel}} D \cdot 1$ and $2 \cdot D =_{\textsc{Rel}} D \cdot 2$ for any D, and thus by P2 this extends to commuting with arbitrary finite languages. For P6, observe that if $C \subseteq_{\textsc{Rel}} D$ then $[\![C \otimes a^*]\!] \in \textsc{Rel}(D)$ for $a \in \mathbb{A}$, which means that $\pi(C) \subseteq \pi(D)$. P7 is a consequence of P4 and the so-called Chrobak normal form for regular languages over unary alphabets [6]. Finally, the proof of P8 is a variant of the proof that the operation of shuffle preserves regularity of languages. ◀

4 Characterization of the Class Containment Problem

Here we give an overview of the ingredients of our decision procedure for class containment.

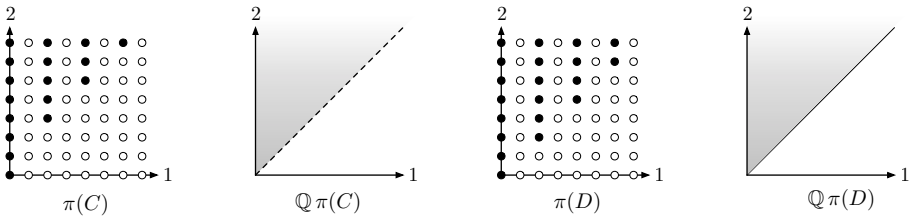
Decomposition. A first ingredient is a decomposition result for regular control languages into $=_{\text{REL}}$-equivalent finite unions of simple languages:

▶ **Proposition 3.** *Every regular language $C \subseteq 2^*$ is effectively $=_{\text{REL}}$-equivalent to a finite union of simple languages.*

To prove the above decomposition result, one begins by applying Lemma 1, so as to decompose the regular language C into a finite union of concat-star languages. Then, each concat-star language is decomposed into unions of concat-star languages of star-height 1 obtaining, for example, $(112(12)^* \cup 122)^* =_{\text{REL}} (122)^* \cup (112 \cup 122)^*112 \cup (112 \cup 122)^*11122 \cup (112 \cup 122)^*1111222$. This latter step is non-trivial to show, and exploits the increased flexibility of the relation $=_{\text{REL}}$ compared to equality. It also exploits in a crucial way properties of linear sets, and more specifically those that result from taking the Parikh images of concat-star languages. Finally, to get the desired decomposition, one needs to decompose further the concat-star languages of star-height 1 into finite unions of simple languages as in, for example, $(12)^*1^*2^* =_{\text{REL}} (12)^*1^* \cup (12)^*2^*$. This last decomposition makes use of some basic properties from Lemma 2.

Parikh ratios. The **Parikh ratio** of a pair $\bar{x} = (n_1, n_2) \in \mathbb{N}^2 \setminus \{(0,0)\}$ is $\rho(\bar{x}) = \frac{n_1}{n_1+n_2}$. We naturally extend this to non-empty words $w \in 2^*$ by letting $\rho(w) = \rho(\pi(w))$ (this describes the proportion of 1's in w). We further extend the notation to languages: $\rho(C) = \{\rho(w) \mid w \in C \setminus \{\varepsilon\}\}$. Note that $\rho(C) \subseteq [0,1]_{\mathbb{Q}}$. It is sometimes useful to think of $\rho(C)$ as the cone $\mathbb{Q}\,\pi(C) = \{q \cdot \pi(w) \mid q \in \mathbb{Q}, w \in C\}$ inside the rational plane $\mathbb{Q} \times \mathbb{Q}$.

▶ **Example 4.** The Parikh images of the languages $C = (2(2112)^*)^*$ and $D = (2 \cup 2112)^*$ are depicted below. Note that $\rho(C) = \left[0, \frac{1}{2}\right)_{\mathbb{Q}}$, while $\rho(D) = \left[0, \frac{1}{2}\right]_{\mathbb{Q}}$.



The following lemma summarizes the main properties of Parikh ratios that we will need.

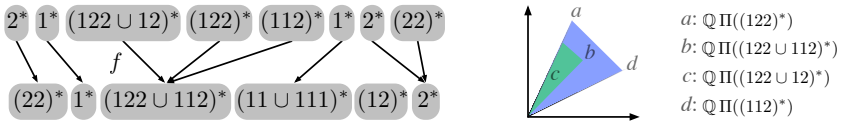
▶ **Lemma 5.** *The Parikh ratio of a concat-star language C verifies the following properties:*
1. *If $C = C_1^* u_1 \cdots C_n^* u_n$, then $\rho(C) \subseteq [\min_i \inf \rho(C_i^*), \max_i \sup \rho(C_i^*)]_{\mathbb{Q}}$;*
2. *Moreover, if $C = D^*$ for a finite D, then $\rho(C) = [\min \rho(D), \max \rho(D)]_{\mathbb{Q}}$.*

Synchronizing morphisms. Another fundamental ingredient is the notion of *synchronizing morphism*, which intuitively relates the components of a concat-star language C to the components of a concat-star language D by comparing the Parikh ratios.

Let $C = C_1^* u_1 \cdots C_k^* u_k$ be a heterogeneous concat-star language and $D = D_1^* v_1 \cdots D_h^* v_h$ any concat-star language. We say that a function $f : [1,k] \to [1,h]$ is a **synchronizing morphism** (abbreviated *s.m.*) from C to D if
- it is monotonic: $f(i) \leq f(j)$ whenever $i \leq j$; and
- it preserves Parikh-ratio: for every $i \in [1,k]$, $\rho(C_i^*) \subseteq \rho(D_{f(i)}^*)$.

We write $C \xrightarrow{s.m.} D$ to denote the existence of such synchronizing morphism. By convention, if C is homogeneous we say that there is a synchroninzing morphism from C to D. In particular, $u \xrightarrow{s.m.} v$ for every $u, v \in \mathbb{2}^*$. The sole purpose of this trivial definition on homogeneous concat-star languages is to make the characterizations statements simpler.

▶ **Example 6.** The following function f is a synchronizing morphism:



Observe that synchronizing morphisms are closed under composition and hence $\xrightarrow{s.m.}$ defines a pre-order on concat-star languages.

Class Containment Problem for simple languages. The existence of synchronizing morphism is the key property that characterizes $\subseteq_{\text{Rel}}$ on simple languages:

▶ **Proposition 7.** *For all simple $C, D \subseteq \mathbb{2}^*$, $C \subseteq_{\text{Rel}} D$ iff $\pi(C) \subseteq \pi(D)$ and $C \xrightarrow{s.m.} D$.*

Note that the case of C homogeneous follows from P8. Intuitively, for any C smooth heterogeneous concat-star language of star-height 1, the characterization says that, every regular $L_1 \subseteq C \otimes \mathbb{A}^*$ can be *resynchronized* to some regular $L_2 \subseteq D \otimes \mathbb{A}^*$ that denotes the same relation (*i.e.*, $[\![L_1]\!] = [\![L_2]\!]$) iff $\pi(C) \subseteq \pi(D)$ and for every component of C, there is a component of D that contains its Parikh ratio. Further, the matching between components is monotonic. For example, we have $(12)^*(112)^* \subseteq_{\text{Rel}} (12 \cup 11122)^*(121)^*1^*2^*$, because the Parikh ratios of $(12)^*$ and $(112)^*$ are included in those of $(12 \cup 11122)^*$ and $(121)^*$, respectively. On the other hand, we have $(112)^*(12)^* \not\subseteq_{\text{Rel}} (12 \cup 11122)^*(121)^*1^*2^*$ because in this case there is no motononic matching between components.

The proof of the above characterization will be the theme of Section 6. The proof of the right-to-left direction requires a *normal form* for star languages C^*, with C finite, that shows that for all $u_-, u_+ \in C$ with minimum and maximum Parikh ratio, and for all $p, q > 0$, there is a finite $C' \subseteq C^*$ so that $C^* =_{\text{Rel}} (u_-^p \cup u_+^q)^* \cdot C'$.

Generalization to unions of simple languages. Section 7 concerns the generalization of the characterization to finite unions of simple languages, which cover arbitrary regular languages up to $=_{\text{Rel}}$-equivalence. Thus, the previous characterization for simple languages constitutes the base case of our characterization. In fact, a first generalization of the characterization for unions on the left hand-side holds thanks to the following trivial lemma.

▶ **Lemma 8.** *$C_1 \cup C_2 \subseteq_{\text{Rel}} D$ iff $C_1 \subseteq_{\text{Rel}} D$ and $C_2 \subseteq_{\text{Rel}} D$.*

The characterization turns out to be more involved when we have unions on the right hand-side. In particular, containment in the union does not imply containment in any of the disjuncts. As a simple example, consider $C = (12)^*$, $D_1 = (112 \cup 1122)^*$, and $D_2 = (122 \cup 1122)^*12$. We have $C \subseteq_{\text{Rel}} D_1 \cup D_2$, although $C \not\subseteq_{\text{Rel}} D_1$ and $C \not\subseteq_{\text{Rel}} D_2$. Neither it holds that Parikh image containment together with the existence of s.m. to one of the disjuncts suffices. As an example, for $C' = (12)^*$, $D_1' = (1212)^*$, $D_2' = 1^*2^*$, we have $C' \not\subseteq_{\text{Rel}} D_1' \cup D_2'$ although $\pi(C') \subseteq \pi(D_1' \cup D_2')$ and $C' \xrightarrow{s.m.} D_1'$.

The characterization we provide is inductive on the number of languages that are unioned on the right hand-side. Concretely, for a union of two languages, we will show that $C \subseteq_{\text{Rel}} D_1 \cup D_2$ iff $C \xrightarrow{s.m.} D_i$ for some i and $C \setminus [D_i]_\pi \subseteq_{\text{Rel}} D_{3-i}$, where $[D_i]_\pi$ is the closure of D_i

under permutations, that is, $[D_i]_\pi \stackrel{\text{def}}{=} \{w \in 2^* \mid \pi(w) \in \pi(D_i)\}$. The idea that underlies the proof of the necessity of our characterization is that C can be split into a disjoint union of $C \cap [D_i]_\pi$ and $C \setminus [D_i]_\pi$, in such a way that $C \cap [D_i]_\pi \subseteq_{\text{REL}} D_i$ and $C \setminus [D_i]_\pi \subseteq_{\text{REL}} D_{3-i}$.

For finite unions of simple languages, we have the following characterization:

▶ **Theorem 9.** *For finite unions $C = \bigcup_i C_i$ and $D = \bigcup_j D_j$ of simple languages, the following are equivalent:*

- $C \subseteq_{\text{REL}} D$,
- *$\forall i\ \exists j$ with $C_i \xrightarrow{s.m.} D_j$ and $\pi(C_i) \subseteq \pi(D)$, and in addition, if C_i is heterogeneous, then $C_i \setminus [D_j]_\pi \subseteq_{\text{REL}} \bigcup_{j' \neq j} D_{j'}$.*

Coming back to the previous example, $C = (12)^*$, $D_1 = (112 \cup 1122)^*$, and $D_2 = (122 \cup 1122)^*12$. We have $C \subseteq_{\text{REL}} D_1 \cup D_2$, one can explain $C \subseteq_{\text{REL}} D_1 \cup D_2$ by the fact of having $C \xrightarrow{s.m.} D_1$ and $C \setminus [D_1]_\pi = (1212)^*12 \subseteq_{\text{REL}} D_2$, where the latter containment holds by the fact that $(1212)^*12 \xrightarrow{s.m.} D_2$ and $\pi((1212)^*12) \subseteq \pi(D_2)$.

Note that hidden in the statement of Theorem 9 there's a caveat: in order for it to be sound, $C_i \setminus [D_j]_\pi$ needs to be *regular*. And in fact this is not the case in general: if $C_i = 1^*2^*$ and $D_j = (12)^*$, we get a non-regular language $C_i \setminus [D_j]_\pi = \{1^n 2^m \mid n \neq m\}$. However, provided $C_i \xrightarrow{s.m.} D_j$ for C_i heterogeneous, we show that $C_i \setminus [D_j]_\pi$ is effectively regular (in the sense that an automaton recognizing it can be computed from automata recognizing C_i and D_j). This is a non-trivial fact, and will be proved in Section 5 (Proposition 11).

The second key ingredient is that if $C_i \subseteq_{\text{REL}} D_1 \cup \cdots \cup D_n$, then there must be some j so that $C_i \xrightarrow{s.m.} D_j$. This will be proved in Section 6 (Lemma 14).

5 Decomposition into simple languages

As already mentioned, we start by reducing the Class Containment Problem for arbitrary regular languages to the case of finite unions of simple languages (Proposition 3 below). We do this in two steps. First, we decompose regular languages into finite unions of concat-star languages of star-height 1 (Lemma 12 below). Then, we further decompose the latter languages into finite unions of simple languages (Lemma 13 below).

Unions of star-height 1 languages. Lemma 12 relies on two key results, which are also of independent interest. The first result is a normal form representation of the Parikh image $\pi(C)$ of a concat-star language C. Formally, we say that a linear set $\langle \bar{x}, P \rangle$ is in **normal form** if the elements of P are linearly independent. We extend this notion to semilinear sets by saying that $\langle \bar{x}_1, P_1 \rangle \cup \cdots \cup \langle \bar{x}_n, P_n \rangle$ is in normal form if the vectors in $\bigcup_i P_i$ are linearly independent. Note that if the representation of a semilinear set is in normal form then all its linear sets are in normal form, but the converse does not hold —for example, consider $\langle \bar{0}, \{(2,0)\} \rangle \cup \langle \bar{0}, \{(3,0)\} \rangle$. The following lemma shows that Parikh images of concat-star languages enjoy normal forms.

▶ **Lemma 10.** *For every concat-star language $C = C_1^* u_1 \cdots C_n^* u_n$, there exists a normal form representation of its Parikh image $\pi(C)$. Moreover, the two periods are $\bar{x}_-$ and $\bar{x}_+$ such that $\rho(\bar{x}_-) = \min_i(\inf \rho(C_i^*))$ and $\rho(\bar{x}_+) = \max_i(\sup \rho(C_i^*))$.*

Proof idea. One considers a regular expression for C with possibly nested Kleene stars, and analyses the dependencies between iterations of the various stars. For example, in the expression $((12)^*11)^*(112)^*$, one cannot iterate the word 12 without first iterating at least once the word 11. This analysis is eased by the construction of a suitable forest, whose nodes correspond to iterable factors and the ancestor relation captures the dependency. ◀

It is worth pointing out a few differences with a seemingly similar normal form from [11]. The normal form from the cited work holds for arbitrary regular languages over arbitrary alphabets. Our normal form holds only for concat-star languages over binary alphabets. On the other hand, the normal form from [11] does not guarantee the linear independence of the vectors in the union of all the periods, as we do here instead. The proof of Proposition 11 below strongly relies on such an additional property.

The second important result shows that, under certain conditions, one can intersect a regular language C by a language of the form $[D]_\pi = \pi^{-1}(\pi(D))$, with D concat-star, and obtain a language that is again regular. This result not only enables the decomposition into star-height 1 languages, but will be used also later to formalize a recursive characterization of $\subseteq_{\text{REL}}$ for unions of simple languages (cf. Section 7).

▶ **Proposition 11.** *Given C regular and D concat-star so that $\rho(\mathit{cycles}(C)) \subseteq \rho(\mathit{cycles}(D))$, the languages $C \cap [D]_\pi$ and $C \setminus [D]_\pi$ are effectively regular. If in addition D is of the form $D_1^* u$, then $C \cap [D]_\pi \subseteq_{\text{REL}} D$.*

Proof idea. We exploit the fact that words in $\mathbb{2}^*$ are in bijection with paths inside $\mathbb{N}^2$ that originate in $\bar{0} = (0,0)$ and, furthermore, that words with the same Parikh image correspond to paths with the same endpoints. The claim boils down to considering some word $w \in \mathbb{2}^*$ and at proving that, under suitable hypotheses, the path induced by w can be approximated by a path inside $\pi(D)$ that stays sufficiently close to the former path. The use of Lemma 10 will be crucial here, since it gives a normal form $\bigcup_i \langle \bar{x}_i, P_i \rangle$ for the latter set $\pi(D)$. Intuitively, it implies that the words from $[D]_\pi$ are represented by paths that never get too far from the linear set $\langle \bar{0}, \bigcup_i P_i \rangle$. For example, by pairing this property with the assumption that $\rho(\mathit{cycles}(C)) \subseteq \rho(\mathit{cycles}(D))$, one can show that the path induced by a word $w \in C$ stays close to $\langle \bar{0}, \bigcup_i P_i \rangle$, and hence also to $\pi(D)$. Stronger variants of this property are shown, that take into account the exact displacement of points along the path induced by w from the points in $\pi(D)$. These latter properties are used by suitable automata that recognize the languages $C \cap [D]_\pi$ and $C \setminus [D]_\pi$. ◀

As we explained in the proof sketch, the above proposition relies on the normal form for the semilinear set $\pi(D)$, which in turns relies on the fact that D is concat-star. The proposition does not hold if we replace D with an arbitrary regular language. For instance, consider $C = 1(11)^*2(22)^*$ and $D = (12)^* \cup (11)^*(22)^*$, and observe that $\rho(\mathit{cycles}(C)) = [0,1]_{\mathbb{Q}} = \rho(\mathit{cycles}(D))$, but $C \cap [D]_\pi = \{1(11)^n 2(22)^n \mid n \in \mathbb{N}\}$ is clearly not regular.

Although Proposition 11 is stated in full generality, that is, for every regular language C so that $\rho(\mathit{cycles}(C)) \subseteq \rho(\mathit{cycles}(D))$, in the proof of the decomposition result below we will use it only for a smooth heterogeneous concat-star language C so that $C \xrightarrow{s.m.} D$ (this is sufficient but not necessary for verifying the hypothesis $\rho(\mathit{cycles}(C)) \subseteq \rho(\mathit{cycles}(D))$).

▶ **Lemma 12.** *Every regular $C \subseteq \mathbb{2}^*$ is $=_{\text{REL}}$-equivalent to a finite union $\bigcup_i D_i$ of concat-star languages of star-height 1.*

Towards the proof of the statement, note that, by Lemma 1, C is a finite union of concat-star languages $C_1^* u_1 \cdots C_n^* u_n$. The lemma then follows from applying Claim 1 below to each component D^* of the concat-star languages, and then using P2.

▶ **Claim 1.** *Every regular D^* is $=_{\text{REL}}$-equivalent to a finite union $\bigcup_i D_i^* u_i$, with finite D_i's.*

Proof idea of Claim 1. Since $\pi(D^*)$ is a finite union of linear sets, from the latter we can extract languages of the form $D_i^* u_i$. Then we can decompose D^* as the union of $D^* \cap [D_i^* u_i]_\pi$. From there, the result follows easily from Proposition 11 and P2. ◀

Unions of simple languages. We finally show how to decompose into simple languages.

▶ **Lemma 13.** *Every concat-star $C \subseteq \mathbb{2}^*$ of star-height 1 is $=_{\text{REL}}$-equivalence to a finite union $\bigcup_i C_i$ of simple languages.*

Proof idea. By using the basic properties given in Lemma 2, we can reduce the problem to the case where C is of the form $1^{k*}2^{\hat{k}*}w^*$ for some heterogeneous word w and some natural numbers $k, \hat{k}$. This case is easy to prove by using again those basic properties. ◀

As a corollary of Lemmas 12 and 13, we have our desired result.

▶ **Proposition 3.** Every regular language $C \subseteq \mathbb{2}^*$ is effectively $=_{\text{REL}}$-equivalent to a finite union of simple languages. ◀

6 Simple languages

We prove the characterization result for simple languages, which we recall here.

▶ **Proposition 7.** For all simple $C, D \subseteq \mathbb{2}^*$, $C \subseteq_{\text{REL}} D$ iff $\pi(C) \subseteq \pi(D)$ and $C \xrightarrow{s.m.} D$.

For the left-to-right direction, by P6, $C \subseteq_{\text{REL}} D$ implies $\pi(C) \subseteq \pi(D)$. The proof that $C \subseteq_{\text{REL}} D$ implies $C \xrightarrow{s.m.} D$ is given in a more general setup where D is a finite union of simple languages. This statement will be used in the characterization of the next section.

▶ **Lemma 14.** *For C a simple language and $D = \bigcup_i D_i$ finite union of simple languages, if $C \subseteq_{\text{REL}} D$, then $C \xrightarrow{s.m.} D_i$ for some i. In particular, for C, D simple languages, if $C \subseteq_{\text{REL}} D$, then $C \xrightarrow{s.m.} D$.*

Proof idea. The idea is to construct a relation $R \in \text{REL}(C)$ so that from $R \in \text{REL}(D)$, using suitable pumping arguments, one can extract a synchronizing morphism from C to some D_i. The relation R must depend on both languages C, D, but the underlying alphabet can be fixed and taken binary, say $\mathbb{A} = \{a, b\}$. For example, if C is of the form C_1^* and contains two words u^- and u^+ with minimum and maximum Parikh ratios, and if the automaton for D has a single strongly connected component, then one can define the relation $R = [\![(u^- \otimes a^{|u^-|})^* \cdot (u^+ \otimes b^{|u^+|})^*]\!]$. In this case, $R \in \text{REL}(D)$ would imply $\rho(u^-), \rho(u^+) \in \rho(D)$, and hence $C \xrightarrow{s.m.} D$. This construction can be modified for more general languages C, D, by using words with different Parikh ratios from each component of C and by increasing the number of alternations between these ratios on the basis of the number of components of D. While the construction is more involved in the general case, and in particular needs to include iterations of words which are not necessarily of minimum or maximum Parikh ratios for a component, the intuition remains the same. ◀

▶ **Observation 15.** *The previous Lemma 14 does not hold for arbitrary concat-star languages C. For example, consider $(12)^*1^*2^* =_{\text{REL}} (12)^*1^* \cup (12)^*2^*$, where there is no s.m. from $(12)^*1^*2^*$ to $(12)^*1^*$, nor from $(12)^*1^*2^*$ to $(12)^*2^*$.*

Conversely, to show that the conditions $\pi(C) \subseteq \pi(D)$ and $C \xrightarrow{s.m.} D$ are sufficient to have $C \subseteq_{\text{REL}} D$, where C, D are simple, it is useful to introduce a normal form for languages of the form C^*, with C finite.

▶ **Lemma 16.** *For every $p, q > 0$, finite $C \subseteq \mathbb{2}^*$, and $u_-, u_+ \in C$ so that $\rho(u_-) = \min \rho(C)$ and $\rho(u_+) = \max \rho(C)$, there exists a finite $C' \subseteq C^*$ so that $C^* =_{\text{REL}} (u_-^p \cup u_+^q)^* \cdot C'$.*

In particular, the lemma implies that $C^* =_{\text{Rel}} (u_- \cup u_+)^* \cdot C'$ for some finite $C' \subseteq C^*$ and u_-, u_+ words of C of minimum and maximum ratio. In other words, it just suffices to iterate two words from C and then append tails of bounded length to obtain the class $\text{Rel}(C^*)$. With this in mind, we can easily prove our characterization for simple languages.

Proof idea of Proposition 7. The left-to-right direction follows from P6 and Lemma 14. For the right-to-left direction, if C is homogeneous, the fact that $\pi(C) \subseteq \pi(D)$ yields $C \subseteq_{\text{Rel}} D$ by P8. Otherwise, we assume wlog that C and D are of the form $C_1^* \cdots C_n^* u$ and $D_1^* \cdots D_m^* v$ (by P5). Since every C_i is finite, one can consider words $w_{i,-}, w_{i,+}$ of minimum and maximum Parikh ratio. Using the normal form of Lemma 16, we obtain that $C_i^* \subseteq_{\text{Rel}} D_{f(i)}^* C_i'$ for a finite $C_i' \subseteq C_i^*$. Thus, $C_1^* \cdots C_n^* u \subseteq_{\text{Rel}} D_{j_1}^* C_1' \cdots D_{j_n}^* C_n' u =_{\text{Rel}} D_{j_1}^* \cdots D_{j_n}^* C_1' \cdots C_n' u \subseteq_{\text{Rel}} D_1^* \cdots D_m^* v$. ◀

7 Regular languages

We now prove the characterization theorem for unions of simple languages. Thanks to this theorem and to Proposition 3, we will obtain an effective characterization for arbitrary regular languages, and thus solve the Class Containment Problem in its full generality.

▶ **Theorem 9.** For finite unions $C = \bigcup_i C_i$ and $D = \bigcup_j D_j$ of simple languages, we have $C \subseteq_{\text{Rel}} D$ if and only if $\forall i\ \exists j$ with $C_i \xrightarrow{s.m.} D_j$ and $\pi(C_i) \subseteq \pi(D)$, and in addition, if C_i is heterogeneous, then $C_i \setminus [D_j]_\pi \subseteq_{\text{Rel}} \bigcup_{j' \neq j} D_{j'}$.

Note in particular that the conditions in the characterization of Theorem 9 require that $C_i \setminus [D_j]_\pi$ is regular whenever $C_i \xrightarrow{s.m.} D_j$ and C_i is heterogeneous. This property is verified with the use of Proposition 11 from Section 5. Indeed, $C_i \xrightarrow{s.m.} D_j$ for C_i heterogeneous implies that all components of C_i are mapped to components of D_j. In view of Lemma 5 and the fact that C_i and D_j have star-height 1, this implies that $\rho(\mathit{cycles}(C_i)) \subseteq \rho(\mathit{cycles}(D_j))$, and hence, by Proposition 11, $C_i \setminus [D_j]_\pi$ is regular. We are now ready to prove the theorem:

Proof of Theorem 9. For the top-to-bottom implication, by Lemma 8, we have that $C_i \subseteq_{\text{Rel}} D$ for every i. Containment of Parikh images follows from P6. For any fixed i, if C_i is homogeneous we have $C_i \xrightarrow{s.m.} D_j$ for every j, and if it is smooth heterogeneous, then Lemma 14 yields the existence of some j so that $C_i \xrightarrow{s.m.} D_j$. By Proposition 11, $C_i \setminus [D_j]_\pi$ is regular, and we now prove that $C_i \setminus [D_j]_\pi \subseteq_{\text{Rel}} \bigcup_{j' \neq j} D_{j'}$. Take $R \in \text{Rel}(C_i \setminus [D_j]_\pi)$ and a regular $L \subseteq (C_i \setminus [D_j]_\pi) \otimes \mathbb{A}^*$ so that $[\![L]\!] = R$. Since $C_i \setminus [D_j]_\pi \subseteq C_i$, we have $R \in \text{Rel}(C_i) \subseteq \text{Rel}(D)$, by P1 and hypothesis. Let $L' \subseteq D \otimes \mathbb{A}^*$ be a regular language so that $[\![L']\!] = [\![L]\!] = R$. Since the projection onto $\mathbb{2}$ of L and L' have necessarily the same Parikh image, it follows that $L' \cap (D_j \otimes \mathbb{A}^*) = \emptyset$, and thus that $L' \subseteq (\bigcup_{j' \neq j} D_{j'}) \otimes \mathbb{A}^*$ or, in other words, that $R \in \text{Rel}(\bigcup_{j' \neq j} D_{j'})$.

For the bottom-to-top implication, for C_i homogeneous, $\pi(C_i) \subseteq \pi(D)$ implies $C_i \subseteq_{\text{Rel}} D$ by P8. For C_i heterogeneous, we have $C_i = (C_i \setminus [D_j]_\pi) \cup (C_i \cap [D_j]_\pi)$. By hypothesis plus property P1, $C_i \setminus [D_j]_\pi \subseteq_{\text{Rel}} D$. Then, by Lemma 8, it only remains to check that $C_i \cap [D_j]_\pi \subseteq_{\text{Rel}} D$. Now, by Proposition 11 and Proposition 3, $C_i \cap [D_j]_\pi$ is $=_{\text{Rel}}$-equivalent to a finite union of simple languages $(C_k')_{k \in K}$. Note that $C_k' \subseteq_{\text{Rel}} C_i$ for all $k \in K$. Then, by the left-to-right direction of Proposition 7, we have $C_k' \xrightarrow{s.m.} C_i$ for all k. By composition of synchronizing morphisms, we obtain $C_k' \xrightarrow{s.m.} D_j$ for all $k \in K$. Since we also have that $\pi(C_k') \subseteq \pi(D_j)$, by the right-to-left direction of Proposition 7, we have that $C_k' \subseteq_{\text{Rel}} D_j$ for all $k \in K$. Then, from Lemma 8 it follows that $C_i \subseteq_{\text{Rel}} D_j \subseteq D$. Since this happens for every C_i, again by Lemma 8 the statement follows. ◀

8 Decidability and complexity

We have given a characterization of the pairs C, D of regular languages that satisfy $C \subseteq_{\text{Rel}} D$. We argue that this characterization is effective.

As explained in Section 7, there are three main steps that one need to take for deciding whether $C \subseteq_{\text{Rel}} D$, for two given regular languages C, D: First, one needs to decompose C and D as finite unions $\bigcup_i C_i$ and $\bigcup_j D_j$ of simple languages. This preprocessing relies on two constructions: the computation of the normal form for semilinear sets and the construction of an automaton for $C \cap [D]_\pi$, proving that is regular. A close inspection of these proofs in Section 5 shows that both procedures are effective, and thus so is the decomposition.

Then, based on the characterization of Theorem 9, one has to identify suitable synchronizing morphisms from each C_i to some D_j. This step boils down to checking whether two components $C^*_{i,i'}$ and $D^*_{j,j'}$ of concat-star languages satisfy $\rho(C^*_{i,i'}) \subseteq \rho(D^*_{j,j'})$. Thanks to the insight of Lemma 5, the containment of Parikh ratios and thus the existence of such synchronizing morphism is decidable.

Finally, the third step uses Theorem 9, reducing the problem $\bigcup_i C_i \subseteq_{\text{Rel}} \bigcup_j D_j$ to subproblems of the form $C_i \setminus [D_{j_i}]_\pi \subseteq_{\text{Rel}} \bigcup_{j' \neq j_i} D_{j'}$, which has a smaller union in the right hand-side and thus can be solved recursively (but in principle non-elementary).

The above arguments show that the Class Containment Problem is decidable. Once we know that $C \subseteq_{\text{Rel}} D$ for two given regular languages C, D, it is reasonable to ask whether it is possible to resynchronize any relation from C to D, namely, whether there is an algorithm that transforms any automaton $\mathcal{A}$ recognizing $L \subseteq C \otimes \mathbb{A}^*$ into an automaton $\mathcal{A}'$ recognizing $L' \subseteq D \otimes \mathbb{A}^*$ so that $[\![L']\!] = [\![L]\!]$. A close inspection to our decision procedure for $C \subseteq_{\text{Rel}} D$ gives a positive answer to the question. Indeed, all our proofs are constructive.

We can summarize the above arguments with the following corollary:

▶ **Corollary 17.** *There is a non-elementary algorithm that, given two regular languages $C, D \subseteq \mathbb{2}^*$, decides whether $C \subseteq_{\text{Rel}} D$.*
There is also a non-elementary algorithm that, given an automaton for $L \subseteq C \otimes \mathbb{A}^*$, constructs an automaton for some $L' \subseteq D \otimes \mathbb{A}^*$ so that $[\![L']\!] = [\![L]\!]$, provided $C \subseteq_{\text{Rel}} D$.

9 Discussion

The overall picture we obtain from our results is that $\text{Rel}(C) \subseteq \text{Rel}(D)$ depend on comparing the ratio growth of the two coordinates on the cycles of the transition graph of the automata $\mathcal{A}_C$, $\mathcal{A}_D$ recognizing C, D. Concretely, our reduction into synchronizing morphisms for simple languages can be thought of restricting our attention to cycles $c_1, \ldots, c_n$ of $\mathcal{A}_C$ so that: c_{i+1} is reachable from c_i, and either c_i or c_{i+1} is heterogeneous. Intuitively, $\text{Rel}(C) \subseteq \text{Rel}(D)$ whenever $\pi(C) \subseteq \pi(D)$ and for every sequence of cycles $c_1, \ldots, c_n$ as before, there exists a corresponding sequence of cycles $c'_1, \ldots, c'_n$ in $\mathcal{A}_D$ with the same properties so that c_i and c'_i have the same Parikh ratio for every i.

We also recall (cf. proof of Lemma 14) that our characterization holds for the containment problem $\text{Rel}(C) \subseteq \text{Rel}(D)$, but also for any variant with a fixed alphabet of cardinality at least 2. For the variant with a unary alphabet $\mathbb{A}$, it is easy to see that $\text{Rel}_{\mathbb{A}}(C) \subseteq \text{Rel}_{\mathbb{A}}(D)$ is equivalent to $\pi(C) \subseteq \pi(D)$. As concerns relations of higher arity defined by control languages $C \subseteq \mathbb{k}^* = [1, k]^*$, it is not clear if a similar characterization may hold. For example, the normal form of Lemma 10 does not generalize to control alphabets of more than two letters. Finally, we leave for future work the issue of determining the precise complexity of the Class Containment Problem.

References

1 Jean Berstel. *Transductions and Context-Free Languages*. B. G. Teubner, 1979.

2 Mikołaj Bojańczyk. Transducers with origin information. In *ICALP*, pages 26–37. Springer, 2014.

3 J.R. Büchi. Weak second-order arithmetic and finite automata. *Mathematical Logic Quarterly*, 6(1-6):66–92, 1960.

4 Olivier Carton, Christian Choffrut, and Serge Grigorieff. Decision problems among the main subfamilies of rational relations. *ITA*, 40(2):255–275, 2006.

5 Christian Choffrut. Relations over words and logic: A chronology. *Bulletin of the EATCS*, 89:159–163, 2006.

6 M. Chrobak. Finite automata and unary languages. *Theoretical Computer Science*, 47(3):149–158, 1986.

7 C. C. Elgot and J. E. Mezei. On relations defined by generalized finite automata. *IBM J. Res. Dev.*, 9(1):47–68, January 1965. URL: `http://dx.doi.org/10.1147/rd.91.0047`, `doi:10.1147/rd.91.0047`.

8 Diego Figueira and Leonid Libkin. Path logics for querying graphs: Combining expressiveness and efficiency. In *LICS*, pages 329–340. IEEE Press, 2015. `doi:10.1109/LICS.2015.39`.

9 Diego Figueira and Leonid Libkin. Synchronizing relations on words. *ACM Transactions on Computer Systems*, 57(2):287–318, 2015. `doi:10.1007/s00224-014-9584-2`.

10 Emmanuel Filiot, Ismaël Jecker, Christof Löding, and Sarah Winter. On equivalence and uniformisation problems for finite transducers. In *ICALP*, volume 55 of *LIPIcs*, pages 125:1–125:14. LZI, 2016. `doi:10.4230/LIPIcs.ICALP.2016.125`.

11 Eryk Kopczyński and Anthony Widjaja To. Parikh images of grammars: Complexity and applications. In *LICS*, pages 80–89. IEEE Press, 2010. `doi:10.1109/LICS.2010.21`.

12 Maurice Nivat. Transduction des langages de Chomsky. *Ann. Inst. Fourier*, 18:339–455, 1968.

A Missing proofs to Section 3

▶ **Lemma 2.** For every regular $C, D, C', D' \subseteq 2^*$,
P1. if $C \subseteq D$, then $C \subseteq_{\text{Rel}} D$;
P2. if $C \subseteq_{\text{Rel}} D$ and $C' \subseteq_{\text{Rel}} D'$, then $C \cdot C' \subseteq_{\text{Rel}} D \cdot D'$ and $C \cup C' \subseteq_{\text{Rel}} D \cup D'$;
P3. if $C \subseteq_{\text{Rel}} D$, then $C^* \subseteq_{\text{Rel}} D^*$;
P4. if $C \subseteq 1^*$ and $C' \subseteq 2^*$, then $C \cdot C' =_{\text{Rel}} C' \cdot C$;
P5. if C is finite, then $C \cdot C' =_{\text{Rel}} C' \cdot C$;
P6. if $C \subseteq_{\text{Rel}} D$ then $\pi(C) \subseteq \pi(D)$; moreover, if C is finite, the converse also holds;
P7. if C is homogeneous concat-star, then $C =_{\text{Rel}} \bigcup_{i \in I} 1^{\ell_i *} 1^{k_i} 2^{\hat{\ell}_i *} 2^{\hat{k}_i}$ for a finite I;
P8. if C is homogeneous concat-star, $C \subseteq_{\text{Rel}} D$ if and only if $\pi(C) \subseteq \pi(D)$.

In order to prove Lemma 2, we will first prove the following decomposition properties:

▶ **Lemma A.1.** *For every regular $C, C' \subseteq 2^*$,*
(a) *If $R \in \text{Rel}(C \cdot C')$, then $R = \bigcup_i R_i \cdot R'_i$ for some $R_1, \ldots, R_n \in \text{Rel}(C)$, $R'_1, \ldots, R'_n \in \text{Rel}(C')$.*
(b) *If $R \in \text{Rel}(C \cup C')$, then $R = R_1 \cup R_2$ for some $R_1 \in \text{Rel}(C)$ and $R_2 \in \text{Rel}(C')$.*
(c) *If $R \in \text{Rel}(C^*)$, then $R = \bigcup_{w \in I} R_{w[1]} \cdots R_{w[|w|]}$ for some regular $I \subseteq \{1, \ldots, n\}^*$ and some $R_1, \ldots, R_n \in \text{Rel}(C)$.*

Proof. *Proof of claim (a).* Since $R \in \text{Rel}(C \cdot C')$ there is some regular language $L \subseteq (2 \times \mathbb{A})^*$ controlled by $C \cdot C'$ with $[\![L]\!] = R$. Let $\mathcal{A} = (Q, q_0, \delta, F)$ be the NFA accepting L. For every $w \in L$ there are w_1 C-controlled, w_2 C'-controlled and $q \in Q$ so that $w_1 \cdot w_2 = w$ and there is an accepting run of $\mathcal{A}$ that reads w_1 and reaches to q, and then from q reads w_2 and reaches a final state. We note this last fact as $w_1 \in \mathcal{L}(\mathcal{A}[q_0, \{q\}])$, $w_2 \in \mathcal{L}(\mathcal{A}[q, F])$. For $q \in Q$ let

$$L_1^q = \{w'_1 \mid w'_1 \text{ is } C\text{-controlled and } w'_1 \in \mathcal{L}(\mathcal{A}[q_0, \{q\}])\},$$
$$L_2^q = \{w'_2 \mid w'_2 \text{ is } C'\text{-controlled and } w'_2 \in \mathcal{L}(\mathcal{A}[q, F])\}.$$

It follows that $L_1^q \cdot L_2^q$ is $(C \cdot C')$-controlled for every $q \in Q$, and we have $[\![L_1^q \cdot L_2^q]\!] = [\![L_1^q]\!] \cdot [\![L_2^q]\!] \subseteq R$. Moreover, by construction, $\bigcup_{q \in Q} [\![L_1^q]\!] \cdot [\![L_2^q]\!] = R$. It remains to see that each L_i^q is regular, which is true due to the following facts:

- Let $H_q = \{w \in (2 \times \mathbb{A})^* \mid w \text{ is } C\text{-controlled}\}$, $H'_q = \{w \in (2 \times \mathbb{A})^* \mid w \text{ is } C'\text{-controlled}\}$. This languages are regular by closure under inverse morphisms of regular languages (the morphism being $(i, a) \mapsto i$ for every $i \in 2$, $a \in \mathbb{A}$).
- Let $\tilde{L_1^q}$ be the closure under prefixes of L and $\tilde{L_2^q}$ the closure under suffixes of L, these are regular since regular languages are closed under prefix closure and suffix closure.
- Finally, L_i^q is regular since it is the intersection of two regular languages: $L_1^q = \tilde{L_1^q} \cap H_q$, $L_2^q = \tilde{L_2^q} \cap H'_q$.

Note that the property does not necessary hold with $n = 1$. For example for $C = C' = \{1\}$ and $R = \{(aa, \varepsilon), (bb, \varepsilon)\}$, since whenever $R \subseteq R_1 \cdot R_2$ for some $R_1, R_2 \in \text{Rel}(C)$, we have that $R_1 \cdot R_2$ contains also the pair (ab, ε).

Proof of claim (b). Since $R \in \text{Rel}(C \cup C')$, there is some regular language $L \subseteq (2 \times \mathbb{A})^*$ controlled by $C \cup C'$ with $[\![L]\!] = R$. For every $w \in L$, w is either C-controlled or C'-controlled. Let

$$L_1 = \{w \in L \mid w \text{ is } C\text{-controlled}\},$$
$$L_2 = \{w \in L \mid w \text{ is } C'\text{-controlled}\}.$$

It follows that $L_1 \cup L_2$ is $(C \cup C')$-controlled and we have $R = [\![L_1 \cup L_2]\!] = [\![L_1]\!] \cup [\![L_2]\!]$. It remains to see that each L_i is regular, which is true due to the following facts:

- Let $H = \{w \in (\mathbb{2} \times \mathbb{A})^* \mid w \text{ is } C\text{-controlled}\}$, $H' = \{w \in (\mathbb{2} \times \mathbb{A})^* \mid w \text{ is } C'\text{-controlled}\}$. This languages are regular by closure under inverse morphisms of regular languages (the morphism being $(i, a) \mapsto i$ for every $i \in \mathbb{2}$, $a \in \mathbb{A}$).
- Finally, L_i is regular since it is the intersection of two regular languages: $L_1 = L \cap H$, $L_2 = L \cap H'$.

Proof of claim (c). We are going to use similar arguments to the ones used in the proof of a. Specifically, instead of factorizing a word $w \in L \subseteq (\mathbb{2} \times \mathbb{A})^*$ into an C-controlled prefix and an C'-controlled suffix, we factorize it as $w = w_1 \cdot w_2 \cdots w_\ell$, for some $\ell \in \mathbb{N}$ and some C-controlled words $w_1, w_2, \ldots, w_\ell$, in such a way that there exist a sequence $q_0, \ldots, q_\ell$ of states of $\mathcal{A}$ such that each language $L_{q_{j-1},q_j} = \mathcal{L}(\mathcal{A}[q_{j-1}, q_j]) \cap (C \otimes \mathbb{A}^*)$ contains the factor w_j. Then we have that $[\![\prod_{j=1}^{\ell} L_{q_{j-1},q_j}]\!] = \prod_{j=1}^{\ell} [\![L_{q_{j-1},q_j}]\!] \subseteq R$. Moreover, by construction, $\bigcup_{q_0 \ldots q_\ell \in Q^*} \prod_{j=1}^{\ell} [\![L_{q_{j-1},q_j}]\!] = R$, from which one easily obtains the desired equation by taking $I = \{(q_0, q_1)(q_1, q_2) \cdots (q_{l-1}, q_l) \mid q_l \text{ is a final state}\} \subseteq Q^2$ and $R_{(q,q')} = [\![L_{q,q'}]\!]$. ◀

Proof of Lemma 2. *Proof of claim P1.* It follows immediately from the definitions.

Proof of claim P2. Let $R \in \textsc{Rel}(C \cdot C')$. By Pa there are $R_1, \ldots, R_n \in \textsc{Rel}(C) \subseteq \textsc{Rel}(D)$ and $R'_1, \ldots, R'_n \in \textsc{Rel}(C') \subseteq \textsc{Rel}(D')$ so that $\bigcup_i R_i \cdot R'_i = R$. It is straightforward from the definition of $[\![\ ]\!]$ to verify that $R_i \cdot R'_i \in \textsc{Rel}(D \cdot D')$ for all i. And, from there, we finally get, using similar arguments, that $R = \bigcup_i R_i \cdot R'_i \in \textsc{Rel}(D \cdot D')$. The proof for the union is similar but using (b) instead of (a).

Proof of claim P3. As we use (a) to prove P2, we can use (c) to prove that this item holds. Given $R \in \textsc{Rel}(C^*)$, we know from (c) that there are some relations $R_1, \ldots, R_n \in \textsc{Rel}(C)$ and a regular language $I \subseteq \{1, \ldots, n\}^*$ such that $R = \bigcup_{i \in I} \prod_{j=1,\ldots,|i|} R_{i[j]}$, where $i[j]$ denotes the j-th letter of the word $i \in I$ (in particular, $i[j] \in \{1, \ldots, n\}$). Since $R_1, \ldots, R_n \in \textsc{Rel}(C) \subseteq \textsc{Rel}(D)$, there are some regular languages $L_1, \ldots, L_n \subseteq D \otimes \mathbb{A}^*$ such that $[\![L_1]\!] = R_1, \ldots, [\![L_n]\!] = R_n$. In particular, we derive that

$$R \;=\; \textstyle\bigcup_{i \in I} \prod_{j=1,\ldots,|i|} [\![L_{i[j]}]\!] \;=\; [\![\bigcup_{i \in I} \prod_{j=1,\ldots,|i|} L_{i[j]}]\!] \;=\; [\![I[k/L_k]]\!]$$

where $I[k/L_k]$ denotes the language over $\mathbb{2} \times \mathbb{A}$ that is obtained from I by substituting every letter $k \in \{1, \ldots, n\}$ with the corresponding regular language L_k. From the fact that regular languages are closed under regular substitutions, we get that $I[k/L_k]$ is a regular language too. Finally, it is easy to see that $I[k/L_k]$ is controlled by D^*, and hence $R \in \textsc{Rel}(D^*)$.

Proof of claim P4. Let $\mathcal{A} = (Q, q_0, \delta, F)$ be a NFA so that $\mathcal{L}(\mathcal{A})$ is $(C \cdot C')$-controlled (and thus $[\![\mathcal{L}(\mathcal{A})]\!] \in \textsc{Rel}(C \cdot C')$). Note that

$$\begin{aligned}
\mathcal{L}(\mathcal{A}) &= \bigcup_{q \in Q} L_{q_0,q} \cdot L_{q,F} \quad \text{for} \\
L_{q_0,q} &= \mathcal{L}(\mathcal{A}[q_0, \{q\}]) \cap (C \otimes \mathbb{A}^*), \\
L_{q,F} &= \mathcal{L}(\mathcal{A}[q, F]) \cap (C' \otimes \mathbb{A}^*).
\end{aligned}$$

Since C and C' have disjoint alphabets, we have that

$$\begin{aligned}
[\![\mathcal{L}(\mathcal{A})]\!] &= [\![\bigcup_{q \in Q} L_{q_0,q} \cdot L_{q,F}]\!] \\
&= [\![\bigcup_{q \in Q} L_{q,F} \cdot L_{q_0,q}]\!] \in \textsc{Rel}(C' \cdot C),
\end{aligned}$$

which proves the statement.

Proof of claim P5. First note that, by second assertion of P2, it is enough to prove the result for $C = \{w\}$, a singleton language. Moreover, by the first assertion of the same item, it suffices to prove the result for the cases where w is a single letter (i.e. $w = 1$ or $w = 2$). We will prove it for $w = 1$, the other case is symmetric. We will see that $\textsc{Rel}(\{1\} \cdot C') \subseteq \textsc{Rel}(C' \cdot \{1\})$, the other containment can be proved in a similar way. Let $R \in \textsc{Rel}(\{1\} \cdot C')$ and $\mathcal{A} = (Q, q_0, \delta, F)$ a NFA such that $\mathcal{L}(\mathcal{A})$ is $\{1\} \cdot C'$-controlled and $[\![\mathcal{L}(\mathcal{A})]\!] = R$. Let $Q_1^a = \{q \in Q \mid \exists \text{ a transition } (q_0, (1, a), q_1) \in \delta\}$. Consider an automaton $\mathcal{A}'$ with states $(q, a) \in Q \times \mathbb{A}$ and transitions

$$\delta' = \{((q, a'), (1, a'), (q', a))\}_{a' \in \mathbb{A}, (q,(1,a),q') \in \delta} \cup \{((q, a'), (2, a), (q', a'))\}_{a' \in \mathbb{A}, (q,(2,a),q') \in \delta}.$$

We don't specify its initial or final states because it is not necessary for our proof.

It is easy to check that

$$[\![\bigcup_{\substack{a,a' \in \mathbb{A}, \\ q_f \in F, \\ q_1 \in Q_1^a}} \mathcal{L}(\mathcal{A}'[(q_1, a), \{(q_f, a')\}]) \cdot \{(1, a')\}]\!] = R$$

and that $\bigcup_{\substack{a,a' \in \mathbb{A}, \\ q_f \in F, \\ q_1 \in Q_1^a}} \mathcal{L}(\mathcal{A}'[(q_1, a), \{(q_f, a')\}]) \cdot \{(1, a')\}$ is a regular $C' \cdot \{1\}$-controlled language which concludes the proof.

Proof of claim P6. Let suppose that $\textsc{Rel}(C) \subseteq \textsc{Rel}(D)$. Towards a contradiction, assume that there exists an element $(\alpha, \beta) \in \pi(C) \setminus \pi(D)$. Now consider the one letter alphabet $\mathbb{A} = \{a\}$ and the singleton relation $R = \{(a^\alpha, a^\beta)\}$. It is easy to check that this relation is in $\textsc{Rel}(C) \setminus \textsc{Rel}(D)$ which is a contradiction. Note that the converse does not hold in general, consider for example $C = (12)^*(1^*|2^*)$, $D = 1^*2^*$. It is clear that $\pi(C) \subseteq \pi(D)$ (in fact they are equal) but one can easily prove that $R = \{(u, v) \mid u = v\} \in \textsc{Rel}(C) \setminus \textsc{Rel}(D)$. By P2, to prove that the converse holds for any finite C, it is enough to prove it for $C = \{w\}$ a singleton language. By hypothesis about Parikh images, there exist $v \in D$ such that $\pi(w) = \pi(w')$. Then it suffices to prove that $\{w\} \subseteq_{\textsc{Rel}} \{w'\}$. For $R \in \textsc{Rel}(w)$, consider a NFA $\mathcal{A} = (Q, q_0, \delta, F)$ such that $\mathcal{L}(\mathcal{A})$ is w-controlled and $[\![\mathcal{L}(\mathcal{A})]\!] = R$. Since w and w' have the same amount of 1's and 2's, for every $w \otimes v \in \mathcal{L}(\mathcal{A})$, there exists a shuffle $\tilde{v}$ of v such that $[\![w \otimes v]\!] = [\![w' \otimes \tilde{v}]\!]$. Then we can construct a NFA $\mathcal{A}'$ such that $\mathcal{L}(\mathcal{A}') = \{w' \otimes \tilde{v} \mid$ there exists $w \otimes v \in \mathcal{L}(\mathcal{A})$ s.t. $[\![w \otimes v]\!] = [\![w' \otimes \tilde{v}]\!]\}$. The result then follows immediatly.

Proof of claim P7. Let $C = C_1^* u_1 \cdots C_k^* u_k$ be concat-star language with all components C_i^* homogeneous. By claim P4 we can swap any two consecutive components C_i^* and C_{i+1}^*, with $C_i^* \subseteq 2^*$ and $C_{i+1}^* \subseteq 1^*$, while preserving $=_{\textsc{Rel}}$-equivalence. Iterating this operation results in a language of the form $C_1' \cdot C_2'$, with $C_1' \subseteq 1^*$ and $C_2' \subseteq 2^*$. In particular, C_1' and C_2' are languages over unary alphabets. There is a special normal form for automata over unary alphabets, called Chrobak normal form [6], that implies that all languages over the unary alphabet $\{1\}$, can be written as unions of languages of the form $1^{\ell_i *} 1^{k_i}$, for i ranging over a finite set I, and similarly for regular languages over $\{2\}$. This proves that $C =_{\textsc{Rel}} C_1' \cdot C_2' = \bigcup_{i \in I} 1^{\ell_i *} 1^{k_i} 2^{\hat{\ell}_i *} 2^{\hat{k}_i}$.

Proof of claim P8. The proof is essentially based on the fact that regularity of languages is preserved under shuffles. Suppose that $\pi(C) \subseteq \pi(D)$ and consider a regular language $L \subseteq C \otimes \mathbb{A}^*$. By the previous claim we can assume that $C = \bigcup_{i \in I} 1^{\ell_i *} 1^{k_i} 2^{\hat{\ell}_i *} 2^{\hat{k}_i}$, and hence $L = \bigcup_{i \in I} L_{1,i} \cdot L_{2,i}$, for some regular languages $L_{1,i} \subseteq (1^{\ell_i *} 1^{k_i}) \otimes \mathbb{A}^*$ and $L_{2,i} \subseteq (2^{\hat{\ell}_i *} 2^{\hat{k}_i}) \otimes \mathbb{A}^*$.

For each $b \in \{1,2\}$ and $i \in I$, let $\mathcal{A}_{b,i}$ be an automaton recognizing $L_{b,i}$, and assume that the states spaces of the $\mathcal{A}_{b,i}$ are pairwise disjoint. Further let $\mathcal{D}$ be an automaton recognizing D. We construct an automaton $\mathcal{B}$ that recognizes $L' \subseteq D \otimes \mathbb{A}^*$ so that $[\![L']\!] = [\![L]\!]$. The automaton $\mathcal{B}$ has for states the triples of states $\mathcal{A}_{1,i}$, $\mathcal{A}_{2,i}$ and $\mathcal{D}$, for all i. When $\mathcal{B}$ is in a control state of the form (q_1, q_2, r), with q_1 state of $\mathcal{A}_{1,i}$ and q_2 state of $\mathcal{A}_{2,i}$, and it reads a letter $(b,a) \in \mathbb{2} \times \mathbb{A}$, it simulates a transition of $\mathcal{D}$ on b, which moves from r to r', and simultaneously a transition of $\mathcal{A}_{b,i}$ on (b,a), which moves from q_i to q_i', and accordingly updates the control state to (q_1', q_2', r'), where $q'_{3-b} = q_{3-b}$. Because $\pi(C) \subseteq \pi(D)$, we know that $\mathcal{B}$ accepts all and only the words $w \in D$ so that $w \in \mathit{shuffle}\{w_1, w_2\}$ for some w_1, w_2, i so that $w_1 \in L_{1,i}$ and $w_2 \in L_{2,i}$. This proves that $[\![w]\!] = [\![w_1 w_2]\!]$. ◀

▶ **Lemma 5.** The Parikh ratio of a concat-star language C verifies the following properties:
1. If $C = C_1^* u_1 \cdots C_n^* u_n$, then $\rho(C) \subseteq [\min_i \inf \rho(C_i^*), \max_i \sup \rho(C_i^*)]_{\mathbb{Q}}$;
2. moreover, if $C = D^*$ for a finite D, then $\rho(C) = [\min \rho(D), \max \rho(D)]_{\mathbb{Q}}$.

Proof. For the first item, note that it is enough to prove that $\min_i(\inf(\rho(C_i^*))) = \inf(\rho(C))$ (and similarly for the maximum). Note that wlog we can assume that C is of the form $C_1^* \cdots C_n^*$ since $\inf \rho(C) = \inf \rho(C_1^* \cdots C_n^*)$. Now, if $C = C_1^* \cdots C_n^*$, every word $w \in C$ has Parikh image $\sum_i \pi(w_i)$ for some words $w_i \in C_i^*$; therefore $\rho(w) \geq \min_i(\inf(\rho(C_i^*)))$. It follows then that $\inf(\rho(C)) \geq \min_i(\inf(\rho(C_i^*)))$. The other inequality follows from the fact that $C_i^* \subseteq C$ for all i.

For the second item, consider an arbitrary rational number $r \in [\min \rho(D), \max \rho(D)]_{\mathbb{Q}}$. For convenience, we fix two words w^- and w^+ in D so that $\rho(w^-) = \min \rho(D)$ and $\rho(w^+) = \min \rho(D)$. Since $r = \rho((w^-)^i \cdot (w^+)^j)$ for some i, j, and $C = D^*$, we have that $r \in \rho(C)$. ◀

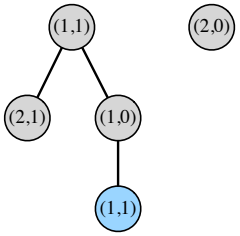


Figure 1 Forest corresponding to $(((12)^*1)^*(112)^*12)^*(11)^*$.

B Missing proofs to Section 5

▶ **Lemma 8.** $C_1 \cup C_2 \subseteq_{\mathrm{REL}} C$ iff $C_1 \subseteq_{\mathrm{REL}} C$ and $C_2 \subseteq_{\mathrm{REL}} C$.

Proof. The left-to-right direction is immedate by transitivity of $\subseteq_{\mathrm{REL}}$ and language inclusion: $C_i \subseteq_{\mathrm{REL}} C_1 \cup C_2$ for $i = 1, 2$. The right-to-left direction is a particular case of P2. ◀

▶ **Lemma 10.** For every concat-star language $C = C_1^* u_1 \cdots C_n^* u_n$, there exists a normal form representation of its Parikh image $\pi(C)$. Moreover, the two periods are $\bar{x}_-$ and $\bar{x}^+$ such that $\rho(\bar{x}^-) = \min_i(\inf \rho(C_i^*))$ and $\rho(\bar{x}^+) = \max_i(\sup \rho(C_i^*))$.

For the proof of Lemma 10, we introduce the notion of *forest-like regular expression*, and we show that for every concat-star regular expression there is a forest-like expression with 'almost' the same Parikh image and finally, we reduce the proof of our lemma to the case of forest-like expressions. We say that a regular expression is **forest-like** if it is defined by the following grammar:

$$F :\stackrel{\text{def}}{=} \varepsilon \mid (Fu)^* \mid FF \qquad \text{for } u \neq \varepsilon.$$

We abstract each forest-like expression as a finite forest whose nodes are labelled with vectors from $\mathbb{N}^2$ in the following way:

- The forest associated to ε is the empty forest.
- The forest associated to an expression $(Fu)^*$ is a tree with a root labeled $\pi(u)$ and set of children $\mathcal{F}$, where $\mathcal{F}$ is the forest associated to F.
- The forest associated to an expression F_1F_2 is the disjoint union of the forests $\mathcal{F}_1$ and $\mathcal{F}_2$ associated to F_1 and F_2 respectively.

Note that, by definition of forest-like, no node of the forest carries a label $(0, 0)$.

▶ **Example B.1.** For the forest-like expression

$$F = (((12)^*1)^*(112)^*12)^*(11)^*$$

we obtain the forest depicted in Figure 1. Intuitively, the ancestor relation represents the dependence between the non-empty words involved in F. Concretely, if a node corresponding to an occurrence of a word u is an ancestor of a node corresponding to an occurrence of a word v, then the regular expression F does not allow any iteration of that occurrence of v without iterating at least once that occurrence of u (have in mind that the same word could occur many times in the expression). In Figure 1 , the node labeled with $(1, 0)$ corresponding to the word 1 in F will be an ancestor of the blue node labeled with $(1, 1)$ corresponding with the first occurrence of 12 in F.

▶ **Lemma B.2.** *Given a concat-star regular expression $C = C_1^* u_1 \cdots C_n^* u_n$, there exists a forest-like regular expression F such that $\pi(C) = \pi(F u_1 \cdots u_n)$.*

Proof. We use the following simple facts about Parikh images. For all regular expression $D_1, \ldots, D_n, D'_1, D'_2$ and words $v_1, \ldots, v_n$ we have:

1. $\pi((\bigcup_{i=1}^n D_i)^*) = \pi(D_1^* \cdots D_n^*)$;
2. $\pi(D_1^* v_1 \cdots D_n^* v_n) = \pi(D_1^* \cdots D_n^* v_1 \cdots v_n)$;
3. if $\pi(D_1) = \pi(D'_1)$ and $\pi(D_2) = \pi(D'_2)$, then $\pi(D_1 D_2) = \pi(D'_1 D'_2)$;
4. if $\pi(D_1) = \pi(D'_1)$, then $\pi((D_1)^*) = \pi((D'_1)^*)$; and
5. $\pi(D_1 \varepsilon) = \pi(D_1)$.

We now prove the statement by induction on the star-height of C.

- Base case: If C has star height 0, then $C = \{\varepsilon\}$ and we take the forest-like expression $F = \varepsilon$.
- Inductive step: Suppose that C has star-height $s > 0$. Then for every $i = 1, \ldots, n$ we have $C_i = \bigcup_{j \in I_i} C_{i,j}$ for $|I_i| < \infty$ and $C_{i,j}$ concat-star expressions of star-height strictly smaller than s for all $j \in I_i$. Then, by inductive hypothesis, we have forest-like expressions $F_{i,j}$ such that $\pi(F_{i,j} u_{i,j}) = \pi(C_{i,j})$ for some words $u_{i,j}$ for all $i = 1, \ldots, n$ and $j \in I_i$. We can take $F = E_{1,1} \cdots E_{|I_1|,1} \cdots E_{1,n} \cdots E_{|I_n|,n}$ where

$$E_{i,j} = \begin{cases} F_{i,j} & \text{if } u_{i,j} = \varepsilon \\ (F_{i,j} u_{i,j})^* & \text{otherwise} \end{cases}.$$

  The result then follows immediately from using fact 2 to move the u_i's to the end, then facts 1 and 3 combined to split the unions inside each C_i, then facts 4 and 3 combined to apply the inductive hypothesis and finally facts 3, 4 and 5 combined to get rid of possible empty words and useless nesting of stars. ◀

▶ **Corollary B.3.** *For any expression of the form $C = C_1^* \cdots C_n^*$, and the forest-like expression F given by Lemma B.2, the forest associated to F has nodes i^-, i^+ labelled $\bar{x}_{i^-}$ and $\bar{x}_{i^+}$ respectively, so that $\rho(\bar{x}_{i^-}) = \min_i(\inf(\rho(C_i^*)))$ and $\rho(\bar{x}_{i^+}) = \max_i(\sup(\rho(C_i^*)))$.*

Proof. Note that it is enough to prove that

$$\min_{j=1,\ldots,n}(\inf(\rho(C_j^*))) =_{(1)} \inf(\rho(C)) =_{(2)} \inf(\rho(F)) =_{(3)} \min_{i \in I} \rho(\bar{x}_i)$$

where I is the set of nodes of the forest and $\bar{x}_i$ is the label of $i \in I$ (the case of max is analogous).

(1) Since $C = C_1^* \cdots C_n^*$, every word $w \in C$ has Parikh image $\sum_{j=1}^n \pi(w_j)$ for some words $w_j \in C_j^*$; therefore $\rho(w) \geq \min_{j=1,\ldots,n}(\inf(\rho(C_j^*)))$. It follows then that $\inf(\rho(C)) \geq \min_{j=1,\ldots,n}(\inf(\rho(C_j^*)))$. The other inequality follows from the fact that $C_j^* \subseteq C$ for all $j = 1, \ldots, n$.

(2) It is immediate from the fact that $\pi(C) = \pi(F)$ (Lemma B.2).

(3) By the grammar defining forest like expressions, every word w in F has Parikh image $\sum_{i \in I} m_i \bar{x}_i$ for some natural numbers m_i; therefore $\rho(w) \geq \min_{i \in I} \rho(\bar{x}_i)$. It follows then that $\inf(\rho(F)) \geq \min_{i \in I} \rho(\bar{x}_i)$. For the other inequality, let $i_0 \in I$ be such that $\rho(\bar{x}_{i_0}) = \min_{i \in I} \rho(\bar{x}_i)$. Now observe that for all $m \in \mathbb{N}$ we can construct words $w_m \in F$ such that $\lim_{m \to \infty} \rho(w_m) = \rho(\bar{x}_{i_0})$ and $\rho(w_m) \geq \rho(\bar{x}_{i_0})$ for all $m \in \mathbb{N}$ (for example the ones that correspond to iterate a word with Parikh image $\bar{x}_{i_0}$ m times and all the words corresponding to the ancestors of the node labeled $\bar{x}_{i_0}$ once). Then $\inf(\rho(F)) \leq \min_{i \in I} \rho(\bar{x}_i)$. ◀

Proof of Lemma 10. We assume, wlog, that $u_1 = \cdots = u_n = \varepsilon$. Indeed, note that if we have a normal form representation $\bigcup_i \langle \bar{x}_i, P_i \rangle$ of $\pi(C_1^* \cdots C_n^*)$, it follows that we have the normal form $\bigcup_i \langle \bar{y} + \bar{x}_i, P_i \rangle$ for $\pi(C)$, where $\bar{y} = \pi(u_1 \cdots u_n)$.

Since we are only interested in the Parikh image of C, by Lemma B.2, wlog we can assume that we have a forest-like representation of C. Let I be the set of nodes of the corresponding forest and for all $i \in I$, let $\bar{x}_i$ be the label of the node i. Let also $\prec \subseteq I \times I$ be the ancestor relation.

Let's fix the following notation:

- $\min_i(\inf(\rho(C_i^*))) = \frac{c^-}{d^-}$ with c^-, d^- coprime, $\max_i(\sup(\rho(C_i^*))) = \frac{c^+}{d^+}$ with c^+, d^+ coprime.
- $I^- = \{i \in I \mid \rho(\bar{x}_i) = \min_i(\inf(\rho(C_i^*)))\}$, $I^+ = \{i \in I \mid \rho(\bar{x}_i) = \max_i(\sup(\rho(C_i^*)))\}$,
- $k^- = mcm(k_i \mid i \in I^-)$ with k_i such that $\bar{x}_i = k_i(c^-, d^- - c^-)$, $k^+ = mcm(k_i \mid i \in I^+)$ with k_i such that $\bar{x}_i = k_i(c^+, d^+ - c^+)$.

Then wlog we can assume (it doesn't change the Parikh image) that for every $i \in I^-$, i has a sibling with label $k^-(c^-, d^- - c^-)$, and similarly for I^+.

Let $\bar{x}^- = k^-(c^-, d^- - c^-)$ and $\bar{x}^+ = k^+(c^+, d^+ - c^+)$. Then, for each $i \in I$, by Lemma C.1, there exist $l_i > 0$, $j_i \geq 0$, $k_i \geq 0$ (j_i, k_i not both equal to 0) such that $l_i\bar{x}_i = j_i\bar{x}^- + k_i\bar{x}^+$. Moreover, we can take $j_i = 0$ if and only if $i \in I^-$ and $k_i = 0$ if and only if $i \in I^+$.

Let $i^-, i^+ \in I$ such that $\bar{x}_{i^-} = \bar{x}^-$, $\bar{x}_{i^+} = \bar{x}^+$, and $N_0 = \max(\{(\sum_{s \prec i^-} j_s + \sum_{s \prec i^+} j_s)\frac{l_i}{j_i} + l_i - 1\}_{i \in I \setminus I^+}, \{(\sum_{s \prec i^-} k_s + \sum_{s \prec i^+} k_s)\frac{l_i}{k_i} + l_i - 1\}_{i \in I \setminus I^-}, \{l_i\}_{i \in I})$. Now consider the following linear sets that we split into four types:

- Type A: $\langle \sum_{i \in I_0} a_i\bar{x}_i, \{\bar{x}^-, \bar{x}^+\} \rangle$ for a set $I_0 \subseteq I$ and $(a_i)_{i \in I_0}$ such that
  - for all $i \in I_0$, $0 < a_i \leq N_0$,
  - I_0 is closed under $\prec$,
  - there exists $i \in I^-$ such that every ancestor of i belongs to I_0 and
  - there exists $i \in I^+$ such that every ancestor of i belongs to I_0.
- Type B: $\langle \sum_{i \in I_0} a_i\bar{x}_i, \{\bar{x}^-\} \rangle$ for a set $I_0 \subseteq I$ such that
  - for all $i \in I_0$, $0 < a_i \leq N_0$,
  - I_0 is closed under $\prec$ and
  - there exists $i \in I^-$ such that every ancestor of i belongs to I_0
- Type C: $\langle \sum_{i \in I_0} a_i\bar{x}_i, \{\bar{x}^+\} \rangle$ for a set $I_0 \subseteq I$ such that
  - for all $i \in I_0$, $0 < a_i \leq N_0$,
  - I_0 is closed under $\prec$ and
  - there exists $i \in I^+$ such that every ancestor of i belongs to I_0
- Type D: $\langle \sum_{i \in I_0} a_i\bar{x}_i, \emptyset \rangle$ for a set $I_0 \subseteq I$ such that
  - for all $i \in I_0$, $0 < a_i \leq N_0$ and
  - I_0 is closed under $\prec$.

It is straightforward to check that the union of all these linear sets is included in $\pi(C)$.

Now we are going to prove that $\pi(C)$ is included in the union. Let $w \in C$, then there exists $\tilde{I}_0 \subseteq I$ closed under $\prec$ such that $\pi(w) = \sum_{i \in \tilde{I}_0} m_i\bar{x}_i$ with $m_i \neq 0$ for all $i \in \tilde{I}_0$. We are going to split the proof in four cases according to the possibilities of having or not iterated at least one word with minimum or maximum Parikh ratio. Intuitively, if we have iterated at least once a word with minimum Parikh ratio, then we are "allowed" to iterate one with Parikh image $\bar{x}^-$ as many times as we want. If not, maybe we didn't need it because we only have iterated a bounded amount of times the cycles that "involve" it. If we want to iterate them many times, then they should be at least as many as we need to allow us to iterate a word with Parikh image $\bar{x}^-$ (see the bound N_0 above).

- Case I: $I^- \cap \tilde{I}_0 \neq \emptyset$ and $I^+ \cap \tilde{I}_0 \neq \emptyset$. Then note that there exist $i \in I^-$ such that all its

ancestors belong to $\tilde{I}_0$ (and similarly for I^+). Then it is easy to check that $\pi(w)$ belongs to the type A linear set given by

$$I_0 \stackrel{\text{def}}{=} \tilde{I}_0 \text{ and } a_i \stackrel{\text{def}}{=} \begin{cases} m_i \bmod l_i & \text{if } m_i \not\equiv 0 \ (l_i) \\ l_i & \text{otherwise.} \end{cases}$$

- Case II: $I^- \cap \tilde{I}_0 \neq \emptyset$, $I^+ \cap \tilde{I}_0 = \emptyset$. Then there exists $i \in I^-$ such that all its ancestors belong to $\tilde{I}_0$. If for all $i \in \tilde{I}_0 \setminus I^-$, $m_i \leq N_0$, then it is easy to check that $\pi(w)$ belongs to the type B linear set given by

$$I_0 \stackrel{\text{def}}{=} \tilde{I}_0 \text{ and } a_i \stackrel{\text{def}}{=} \begin{cases} m_i & \text{if } i \in \tilde{I}_0 \setminus I^- \\ m_i \bmod l_i & \text{if } i \in \tilde{I}_0 \cap I^- \text{ and } m_i \not\equiv 0 \ (l_i) \\ l_i & \text{otherwise.} \end{cases}$$

  If there exist $i_0 \in \tilde{I}_0 \setminus I^-$ such that $m_{i_0} > N_0$, then $\lfloor \frac{m_{i_0}}{l_{i_0}} \rfloor > \frac{N_0+1-l_{i_0}}{l_{i_0}}$ and so, by definition of N_0 plus the fact that $i_0 \notin I^- \cup I^+$, $(\lfloor \frac{m_{i_0}}{l_{i_0}} \rfloor - 1) j_{i_0} > \sum_{s \prec i^+} j_s$ and $(\lfloor \frac{m_{i_0}}{l_{i_0}} \rfloor - 1) k_{i_0} > \sum_{s \prec i^+} k_s$. Then, it is not difficult to check that $\pi(w)$ belongs to the type A linear set given by $I_0 \stackrel{\text{def}}{=} \tilde{I}_0 \cup \{s \in I \mid s \prec i^+\}$ and

$$a_i \stackrel{\text{def}}{=} \begin{cases} m_i \bmod l_i & \text{if } i \in \tilde{I}_0 \text{ and } m_i \not\equiv 0 \ (l_i) \\ l_i & \text{if } m_i \equiv 0 \ (l_i) \text{ or } i \in \{s \in I \mid s \prec i^+\} \setminus \tilde{I}_0. \end{cases}$$

- Case III: $I^- \cap \tilde{I}_0 = \emptyset$, $I^+ \cap \tilde{I}_0 \neq \emptyset$. It is symmetric to the previous case.
- Case IV: $I^- \cap \tilde{I}_0 = \emptyset$, $I^+ \cap \tilde{I}_0 = \emptyset$. If for all $i \in \tilde{I}_0$, $m_i \leq N_0$, then $\pi(w)$ belongs to the type D linear set given by $I_0 \stackrel{\text{def}}{=} \tilde{I}_0$ and $a_i \stackrel{\text{def}}{=} m_i$ for all $i \in \tilde{I}_0$. If there exist $i_0 \in \tilde{I}_0$ such that $m_{i_0} > N_0$, then $\lfloor \frac{m_{i_0}}{l_{i_0}} \rfloor > \frac{N_0+1-l_{i_0}}{l_{i_0}}$ and so, by definition of N_0 plus the fact that $i_0 \notin I^- \cup I^+$, $(\lfloor \frac{m_{i_0}}{l_{i_0}} \rfloor - 1) j_{i_0} > \sum_{s \prec i^-} j_s + \sum_{s \prec i^+} j_s$ and $(\lfloor \frac{m_{i_0}}{l_{i_0}} \rfloor - 1) k_{i_0} > \sum_{s \prec i^-} k_s + \sum_{s \prec i^+} k_s$. Then, it is not difficult to check that $\pi(w)$ belongs to the type A linear set given by $I_0 \stackrel{\text{def}}{=} \tilde{I}_0 \cup \{s \in I \mid s \prec i^-\} \cup \{s \in I \mid s \prec i^+\}$ and

$$a_i \stackrel{\text{def}}{=} \begin{cases} m_i \bmod l_i & \text{if } i \in \tilde{I}_0 \text{ and } m_i \not\equiv 0 \ (l_i) \\ l_i & \text{if } m_i \equiv 0 \ (l_i) \text{ or } i \in (\{s \in I \mid s \prec i^-\} \cup \{s \in I \mid s \prec i^+\}) \setminus \tilde{I}_0. \end{cases}$$

◀

▶ **Proposition 11.** Given C regular and D concat-star so that $\rho(\mathit{cycles}(C)) \subseteq \rho(\mathit{cycles}(D))$, the languages $C \cap [D]_\pi$ and $C \setminus [D]_\pi$ are effectively regular. If in addition D is of the form $D_1^* u$, then $C \cap [D]_\pi \subseteq_{\text{REL}} D$.

For the proof of Proposition 11, we need some auxiliary notation and lemmas.

We will measure the **distance** of a vector $\bar{x} \in \mathbb{N}^2$ from a given set $S \subseteq \mathbb{N}^2$ of vectors by using the ∞-norm, namely, by letting $\mathit{dist}(\bar{x}, S) \stackrel{\text{def}}{=} \min\{\|\bar{x} - \bar{y}\|_\infty \ : \ \bar{y} \in S\}$.

▶ **Lemma B.4.** *Let L be the language recognized by a minimal deterministic automaton with k states and let $\mathit{simple\text{-}cycles}(L) = \{w_1, \ldots, w_n\}$. For every factor w of a word in L, there exists a word w_0 with $|w_0| \leq k$ and $k_1, \ldots, k_n \in \mathbb{N}$ so that $w \in \mathit{shuffle}\{w_0, w_1^{k_1}, \ldots, w_n^{k_n}\}$.*

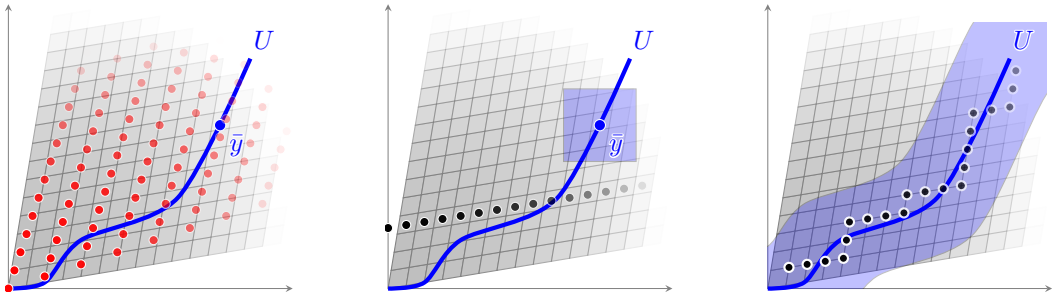


Figure 2 Distance of $\bar{y} \in U$ from the linear sets $\langle \bar{0}, P\rangle$ and $\langle \bar{x}, P'\rangle$, with $P' = \{\bar{x}_+\}$.

Proof. Let $\mathcal{A}$ be an automaton with k states that recognizes L. By way of contradiction, suppose that w is a word of *minimal* length that violates the claim. In particular, w is a factor of some word in L and its length exceeds k (otherwise, $w \in \mathit{shuffle}\{w_0\}$ for $w_0 = w$). Since w is a factor of a word in L, there exist u, v and a successful run $\gamma_1\,\gamma_2\,\gamma_3$ of $\mathcal{A}$, where γ_1 reads u, γ_2 reads w, and γ_3 reads v. Since $|w| > k$, γ_2 can be further decomposed into $\gamma_1'\,\gamma_2'\,\gamma_3'$, with γ_2' *simple cycle*. Let u', w', v' be the words read by $\gamma_1', \gamma_2', \gamma_3'$, respectively, so that $w = u' \cdot w' \cdot v'$. Since γ_2' is a simple cycle, $w' = w_i$ for some $1 \leq i \leq n$, and hence $w \in \mathit{shuffle}\{u'v', w_i\}$. Moreover, since $w' \neq \varepsilon$, $u'v'$ is shorter than w, and hence by the minimality of w, $u'v'$ must satisfy the claim: $u'v' \in \mathit{shuffle}\{w_0, w_1^{k_1}, \ldots, w_n^{k_n}\}$, for some w_0 of length at most k and some $k_1, \ldots, k_n \in \mathbb{N}$. By transitivity we conclude that $w \in \mathit{shuffle}\{w_0, w_1^{k_1'}, \ldots, w_n^{k_n'}\}$ where $k_i' = k_i + 1$ and $k_j' = k_j$ for all $j \neq i$. ◀

▶ **Corollary B.5.** *Let L be the language recognized by a minimal deterministic automaton with k states. For every factor u of a word in L, $\mathit{dist}(\pi(u), \pi(\mathit{simple\text{-}cycles}(L)^*)) \leq k$.*

▶ **Lemma B.6.** *Let $P = \{\bar{x}_-, \bar{x}_+\}$, with $\rho(\bar{x}_-) < \rho(\bar{x}_+)$, let $C \subseteq \mathbb{2}^*$ be regular so that $\rho(\mathit{cycles}(C)) \subseteq [\rho(\bar{x}_-), \rho(\bar{x}_+)]_{\mathbb{Q}}$, and let $\langle \bar{x}, P'\rangle$ be a linear set with $P' \subseteq P$. There is a constant k so that for every prefix w of a word in $C \cap \pi^{-1}(\langle \bar{x}, P'\rangle)$, $\mathit{dist}(\pi(w), \langle \bar{x}, P'\rangle) \leq k$.*

Proof. We will exploit a natural correspondence between words in $\mathbb{2}^*$ and paths inside the discrete plane $\mathbb{N}^2$. Intuitively, each word $u \in \mathbb{2}^*$ induces a path that starts at the origin $\bar{0} = (0, 0)$ of the plane and visits all the vectors corresponding to the Parikh images of prefixes of u. In particular, paths with the same endpoints correspond to words with the same Parikh image. Our first goal is to prove a slightly different claim that concerns the factors of words in C (not necessarily having Parikh image in $\langle \bar{x}, P'\rangle$). We prove that such factors are at bounded distance from the linear set $\langle \bar{0}, P\rangle$:

▶ **Claim.** *For every factor u of a word in C, $\mathit{dist}(\pi(u), \langle \bar{x}, P\rangle) \leq k_1 + k_2$, where k_1 is the number of states of the minimum deterministic automaton for C and $k_2 = \max(\|\bar{x}_-\|_\infty, \|\bar{x}_+\|_\infty)$.*

Proof of claim. We fix a factor u of some word in C, and consider the induced path $U = \{\pi(u') \mid u' \text{ prefix of } u\}$. For this proof the reader can refer to the left hand-side of Figure 2. There, the path U is depicted in blue, and the linear set $\langle \bar{0}, P\rangle$ is represented by a gray grid. Consider a point $\bar{y} \in U$, which corresponds to a prefix u' of u. Since u' is a factor of a word in C, by Corollary B.5 $\bar{y}$ is at distance at most k_1 from $\pi(\mathit{simple\text{-}cycles}(C)^*)$. Figure 2 represents the points in $\pi(\mathit{simple\text{-}cycles}(C)^*)$ by the red dots.

Recall that $\rho(\mathit{cycles}(C)) \subseteq [\rho(\bar{x}_-), \rho(\bar{x}_+)]_{\mathbb{Q}}$. Intuitively, this means that $\pi(\mathit{cycles}(C))$ is contained in a 'cone' with slopes between $\rho(\bar{x}_-)$ and $\rho(\bar{x}_+)$ (this is represented by the gray

shaded area behind the grid). We formalize this notion of 'cone' as follows: given any subset S of the discrete plane, we let $\mathbb{Q}S = \{\alpha\bar{z} \mid \alpha \in \mathbb{Q}, \bar{z} \in S\}$, which is precisely our cone living in the rational plane $\mathbb{Q}^2$. We generalize in the obvious way the distance function to points $\bar{z}$ and subsets $\mathbb{Q}S$ of the rational plane (for this we need to take the infimum instead of the minimum of the distances of $\bar{z}$ from the points of $\mathbb{Q}S$).

From the previous containments, we obtain $\pi(\mathit{cycles}(C)) \subseteq \mathbb{Q}\langle\bar{0}, P\rangle$. Moreover, by construction, every point in $\mathbb{Q}\langle\bar{0}, P\rangle$ is at distance at most $k_2 = \max(\|\bar{x}_-\|_\infty, \|\bar{x}_+\|_\infty)$ from the linear set $\langle\bar{0}, P\rangle$. Since $\pi(\mathit{simple\text{-}cycles}(C)^*) \subseteq \pi(\mathit{cycles}(C))$, we conclude by transitivity that $\bar{y}$ is at distance at most $k_1 + k_2$ from $\langle\bar{0}, P\rangle$. ◀

Now, it remains to extend the previous property to any linear set $\langle\bar{x}, P'\rangle$, with $\bar{x} \in \mathbb{N}^2$ and $P' \subsetneq P$. We restrict our attention to the case $P' = \{\bar{x}_+\}$ (the case $P' = \{\bar{x}_-\}$ is symmetric, the case $P' = \emptyset$ is straightforward, and the case $P' = P$ follows readily from the previous claim). We also use different hypotheses that those of the claim: here we assume that $u \in C$ and its Parikh image belongs to $\langle\bar{x}, P'\rangle$. Under these hypotheses, we prove that the points along the induced path $U = \{\pi(u') \mid u' \text{ prefix of } u\}$ are at distance at most k from $\langle\bar{x}, P'\rangle$, where $k = k_1 + k_2 + k_3$, with k_1, k_2 defined as in the claim and $k_3 = \|\bar{x}\|_\infty$. For this proof the reader may refer to the middle of Figure 2. For example, the black dots represent the set $\langle\bar{x}, P'\rangle$. It will be convenient to have a shorthand notation for denoting neighborhoods of sets of points: given any set $S \subseteq \mathbb{N}^2$, we let $\mathcal{N}_k(S) = \{\bar{x} \in \mathbb{N}^2 \mid \mathit{dist}(\bar{x}, S) \leq k\}$.

Suppose by way of contradiction that there is a point $\bar{y} \in U$ such that $\mathit{dist}(\bar{y}, \langle\bar{x}, P'\rangle) > k = k_1 + k_2 + k_3$. This is equivalent to having $\mathcal{N}_k(\{\bar{y}\}) \cap \langle\bar{x}, P'\rangle = \emptyset$ (this is represented in the figure by a blue square disjoint from the black dots). We first claim that $\mathcal{N}_k(\{\bar{y}\})$ is *above* $\langle\bar{x}, P'\rangle$. Indeed, $\bar{y}$ is the Parikh image of a factor of some word in C, and hence from the previous claim $\mathit{dist}(\bar{y}, \langle\bar{0}, P\rangle) \leq k_1 + k_2$. This is equivalent to saying that $\mathcal{N}_{k_1+k_2}(\{\bar{y}\})$ intersects $\langle\bar{0}, P\rangle$, and hence, since $k = k_1 + k_2 + \|\bar{x}\|_\infty$, $\mathcal{N}_k(\{\bar{y}\})$ intersects $\langle\bar{x}, P\rangle$. However, by our assumption, $\mathcal{N}_k(\{\bar{y}\})$ does not intersect $\langle\bar{x}, P'\rangle$. Since there is no point in $\langle\bar{x}, P\rangle$ that is strictly below $\langle\bar{x}, P'\rangle$, we conclude that $\mathcal{N}_k(\{\bar{y}\})$ must be disjoint and above $\langle\bar{x}, P'\rangle$.

Now, we consider the linear sets $\langle\bar{y}, P\rangle$ and $\langle\bar{y}, P'\rangle$ that originate in the point $\bar{y}$ (these are not shown in the figure). Because $\mathcal{N}_k(\{\bar{y}\})$ is disjoint from $\langle\bar{x}, P'\rangle$, $\mathcal{N}_k(\langle\bar{y}, P'\rangle)$ too is disjoint from $\langle\bar{x}, P'\rangle$. Moreover, because $\mathcal{N}_k(\{\bar{y}\})$ is above $\langle\bar{x}, P'\rangle$, $\mathcal{N}_k(\langle\bar{y}, P'\rangle)$ too is above $\langle\bar{x}, P'\rangle$. This implies that $\mathcal{N}_k(\langle\bar{y}, P\rangle)$ is disjoint from $\langle\bar{x}, P'\rangle$.

Towards a conclusion, consider the vector $\bar{z} = \pi(u) - \bar{y}$, which corresponds to another factor of a word in C. From the usual arguments it follows that $\mathit{dist}(\bar{z}, \langle\bar{0}, P\rangle) \leq k_1 + k_2$, or equally $\bar{z} \in \mathcal{N}_{k_1+k_2}(\langle\bar{0}, P\rangle)$, which implies $\pi(u) \in \mathcal{N}_k(\langle\bar{y}, P\rangle)$. Together with $\mathcal{N}_k(\langle\bar{y}, P\rangle) \cap \langle\bar{x}, P'\rangle = \emptyset$, this implies $\pi(u) \notin \langle\bar{x}, P'\rangle$, which contradicts the initial hypothesis. ◀

We are now ready to prove Proposition 11 (recall that this concerns the regularity of the languages $C \cap [D]_\pi$ and $C \setminus [D]_\pi$).

Proof of Proposition 11. To prove the first claim, it suffices to construct an automaton $\mathcal{D}$ for $C \cap [D]_\pi$ (from this it will follow that the language $C \setminus [D]_\pi$ is also effectively regular, since it can be rewritten as $C \setminus (C \cap [D]_\pi)$ and regular languages are closed under set difference). We fix a word $u \in C \cap [D]_\pi$ and we consider the induced set U of Parikh images of prefixes of u. Let $\langle\bar{x}_1, P_1\rangle \cup \cdots \cup \langle\bar{x}_n, P_n\rangle$ be the normal form of the semilinear set $\pi(D)$ obtained from Lemma 10, where $P_i \subseteq \{\bar{x}_-, \bar{x}_+\}$ for some vectors $\bar{x}_-, \bar{x}_+$ so that $\rho(\bar{x}_-) \leq \rho(\bar{x}_+)$.

Since $\rho(\mathit{cycles}(C)) \subseteq \rho(\mathit{cycles}(D))$, we have $\rho(\mathit{cycles}(C)) \subseteq [\rho(\bar{x}_-), \rho(\bar{x}_+)]_\mathbb{Q}$. Moreover, since $u \in [D]_\pi$, there is an index i so that $u \in \pi^{-1}(\langle\bar{x}_i, P_i\rangle)$. The construction that follows

depends implicitly on the index i, which needs to be correctly guessed by the automaton $\mathcal{D}$ that recognizes $C \cap [D]_\pi$.

By Lemma B.6, we know that the path U can be approximated by another path that is contained in the semilinear set $\langle \bar{x}_i, P_i \rangle$, but that remains also sufficiently close to U (cf. the black dots in the right hand-side of Figure 2). More precisely, there is a constant k that depends only on C and D and so that for every $j = 0, \ldots, |u|$, the following set is non-empty:

$$S_j \stackrel{\text{def}}{=} \langle \bar{x}_i, P_i \rangle \cap \mathcal{N}_k\big(\pi(u[1,j])\big)$$

(here we reuse the notation $\mathcal{N}$ for neighborhoods of sets of points, that was introduced in the proof of Lemma B.6).

The automaton $\mathcal{D}$ exploits the above property to parse the input u while simulating the transitions of an automaton for C, and, at the same time, guessing a series of points $\bar{y}_0 \leq \bar{y}_1 \leq \cdots \leq \bar{y}_{|u|}$ from the sets $S_0, S_1, \ldots, S_{|u|}$. In fact, $\mathcal{D}$ only maintains in memory the **offset** $\bar{z}_j = \bar{y}_j - \pi(u[1,j])$ associated with the current position j — this is possible because $\bar{z}_j$ ranges over the finite domain $[-k,k]^2$. The possible choices for the next offset $\bar{z}_{j+1}$ can be computed knowing the current offset $\bar{z}_j$ and the current input symbol $u[j+1]$. Indeed, we have $\bar{z}_{j+1} - \bar{z}_j = \bar{y}_{j+1} - \bar{y}_j - \pi(u[j+1])$ and $\bar{y}_{j+1} - \bar{y}_j \in \langle \bar{0}, P_i \rangle$, which implies that $\bar{z}_{j+1}$ must range in the set $\big(\bar{z}_j - \pi(u[j+1]) + \langle \bar{0}, P_i \rangle\big) \cap [-k,k]^2$. In the end, the automaton $\mathcal{D}$ accepts the input u if $u \in C$ and the final vector $\bar{z}_{|u|}$ is $\bar{0}$, namely, if $u \in C \cap \pi^{-1}(\langle \bar{x}_i, P_i \rangle)$.

We now turn to the proof of the second claim. We assume, without any loss of generality, that the concact-star language D is of the form $u_0 D_1^*$, since this will simplify notation. Note that nevertheless $u_0 D_1^* =_{\text{REL}} D_1^* u_0$ by P5, and hence we also have that $\pi(u_0 D_1^*) = \pi(D_1^* u_0)$ and $[u_0 D_1^*]_\pi = [D_1^* u_0]_\pi$. Namely, we will show that $C \cap [D]_\pi \subseteq_{\text{REL}} D$ whenever $D = u_0 D_1^*$. The basic idea it to lift the previous construction for the automaton $\mathcal{D}$ to a resynchronization from $(C \cap [D]_\pi) \otimes \mathbb{A}^*$ to $D \otimes \mathbb{A}^*$, namely, to a functional transducer $\mathcal{T}$ that maps words $w \in (C \cap [D]_\pi) \otimes \mathbb{A}^*$ to words $\hat{w} \in D \otimes \mathbb{A}^*$ so that $[\![w]\!] = [\![\hat{w}]\!]$. Intuitively, the transducer $\mathcal{T}$ will use the offsets $\bar{z}_0, \bar{z}_1, \ldots, \bar{z}_{|w|}$ guessed by the previous automaton $\mathcal{D}$ and associate with them a corresponding sequence of outputs $\hat{w}_0, \hat{w}_1, \ldots, \hat{w}_{|w|}$, with $\hat{w}_0 \in u_0 D_1^* \otimes \mathbb{A}^*$ and $\hat{w}_j \in D_1^* \otimes \mathbb{A}^*$ for all $j > 0$. The concatenation of such outputs will give precisely the word $\hat{w} \in D \otimes \mathbb{A}^*$ such that $[\![\hat{w}]\!] = [\![w]\!]$. Once the transducer $\mathcal{T}$ is defined, it could be used to compute from any given regular language $L \subseteq (C \cap [D]_\pi) \otimes \mathbb{A}^*$ a new regular language $\mathcal{T}(L) \subseteq D \otimes \mathbb{A}^*$ so that $[\![L]\!] = [\![L']\!]$, thus showing that $C \cap [D]_\pi \subseteq_{\text{REL}} D$ holds effectively.

To avoid heavy notation, it is convenient to lift the operation of concatenation from words over $\mathbb{A}$ to *pairs* of words over $\mathbb{A}$ in a pointwise manner. Accordingly, we say that a pair p is a prefix of another pair p' if there is a third pair p'' so that $p' = p \cdot p''$; similarly, we say that $p = (w_1, w_2)$ has length $|p| = (n_1, n_2)$ if $|w_1| = n_1$ and $|w_2| = n_2$. Thanks to this, we can succinctly write equations like $[\![w[1,j+1]]\!] = [\![w[1,j]]\!] \cdot [\![w[j+1]]\!]$, or say that $[\![w[1,j]]\!]$ has length $\pi(w[1,j])$ for any $w \in (\mathbb{2} \times \mathbb{A})^*$. We extend further this notation by working with words (or even pairs of words) over the *free group* $\mathbb{A} \cup \mathbb{A}^{-1}$. For example, we could write $(abc) \cdot (bc)^{-1} = (abc) \cdot c^{-1} \cdot b^{-1} = a$. We will tacitly assume that we never construct words with irreducible factors of the form $a \cdot b^{-1}$ with $a \neq b$. Moreover, by a slight abuse of terminology, we say that a word has length $-\ell$ if it is the inverse of a word over $\mathbb{A}$ of length ℓ, and similarly for pairs (note that we do not define the length of words that contain both symbols from $\mathbb{A}$ and symbols from $\mathbb{A}^{-1}$, since these should be first reduced).

As usual, we restrict our attention to an arbitrary word $w = u \otimes v$ with control sequence $u \in C \cap \pi^{-1}(\langle \bar{x}_i, P_i \rangle)$, for some $i = 1, \ldots, n$. We recall that the automaton $\mathcal{D}$ that recognizes $C \cap [D]_\pi$ can guess the same index i, and a series of offsets $\bar{z}_0, \bar{z}_1, \ldots, \bar{z}_{|w|}$ so that, for all

$j = 0, \ldots, |w|$,

$$\bar{y}_j = \bar{z}_j + \pi(u[1,j]) \ \in \ S_j = \langle \bar{x}_i, P_i \rangle \ \cap \ \mathcal{N}_k\big(\pi(u[1,j])\big).$$

The transducer $\mathcal{T}$ will simulate the guesses of $\mathcal{D}$ internally.

Now, let p_j denote the unique prefix of the pair $[\![w]\!]$ that has length $\bar{y}_j$. Since $\bar{y}_j \leq \bar{y}_{j+1}$, each p_j is a prefix of p_{j+1}. We define the **gap** at position $j = 1, \ldots, |w|$, as the difference between two consecutive pairs p_j and p_{j+1}:

$$g_j \ \stackrel{\text{def}}{=} \ (p_j)^{-1} \cdot p_{j+1}.$$

Note that every g_j is a word over $\mathbb{A}$, because p_j is a prefix of p_{j+1}. The goal of the transducer is to produce first a synchronization for the pair p_0, and then some synchronizations for the gaps $g_1, \ldots, g_{|w|}$, so that the total output will be a synchronization of $p_{|w|}$. Under the assumption that the automaton $\mathcal{D}$ for $C \cap \pi^{-1}(\langle \bar{x}_i, P_i \rangle)$ accepts the underlying control sequence u, we will have that $\bar{z}_{|w|} = \bar{0}$ and hence $\bar{y}_{|w|} = \bar{z}_{|w|} + \pi(u) = \pi(u)$. This will imply $p_{|w|} = [\![w]\!]$. Below we explain in more detail how $\mathcal{T}$ can produce the correct outputs.

We introduce a second object, called **lag**:

$$\ell_j \ \stackrel{\text{def}}{=} \ [\![w[1,j]]\!]^{-1} \cdot p_j.$$

Intuitively, ℓ_j is for $\mathcal{T}$ what the offset $\bar{z}_j$ was for $\mathcal{D}$, namely, ℓ_j is the difference between the pair p_j of length $\bar{y}_j$ and the pair encoded by the prefix $w[1,j]$ of the input. Observe that $\ell_j \in (\mathbb{A}^* \cup (\mathbb{A}^{-1})^*) \times (\mathbb{A}^* \cup (\mathbb{A}^{-1})^*)$. This means that ℓ_j may contain words over $\mathbb{A}^{-1}$, even if it is maximally reduced. For example, if p_j is a prefix of $[\![w[1,j]]\!]$, then the lag ℓ_j consists of a pair of words over $\mathbb{A}^{-1}$, meaning that both coordinates of p_j are 'lagging behind' the coordinates of $[\![w[1,j]]\!]$. In general, each coordinate of p_j can lag behind or ahead of the corresponding coordinate of $[\![w[1,j]]\!]$.

We observe that $|\ell_j| = \bar{y}_j - \pi(u[1,j]) = \bar{z}_j \in [-k,k]^2$, which means that lags can be maintained by $\mathcal{T}$ using finitely many states. More precisely, at each position j, $\mathcal{T}$ guesses ℓ_j as any pair of words over $(\mathbb{A}^* \cup (\mathbb{A}^{-1})^*) \times (\mathbb{A}^* \cup (\mathbb{A}^{-1})^*)$ of length $\bar{z}_j$ (the latter vector is available from the underlying automaton $\mathcal{D}$). The correctness of the guesses for the lags ℓ_j are verified implicitly when performing concatenations: if at any moment a concatenation with ℓ_j or its inverse induces an irreducible factor ab^{-1}, with $a \neq b$, then the computation fails, meaning that some words were wrongly guessed.

We now relate the lags to the gaps:

$$\begin{aligned}
g_j \ &= \ (p_j)^{-1} \cdot p_{j+1} && \text{(by definition of gap)} \\
&= \ \big([\![w[1,j]]\!]^{-1} \cdot p_j\big)^{-1} \cdot [\![w[j+1]]\!] \cdot \big([\![w[1,j+1]]\!]^{-1} \cdot p_{j+1}\big) && \text{(by definition of lag)} \\
&= \ (\ell_j)^{-1} \cdot [\![w[j+1]]\!] \cdot \ell_{j+1}. && \text{(by reducing factors)}
\end{aligned}$$

Intuitively, the gap can be equally seen as the difference between two consecutive lags, adjusted by taking into account the current input letter $w[j+1]$. The above property shows that $\mathcal{T}$ can compute g_j using only the bounded amount of information relative to ℓ_j, ℓ_{j+1}, and $w[j+1]$.

Summing up, the transducer $\mathcal{T}$ parses the input w, while guessing some offsets $\bar{z}_0, \ldots, \bar{z}_{|w|}$ and some lags $\ell_0, \ldots, \ell_{|w|}$, and producing some outputs $\hat{w}_0, \ldots, \hat{w}_{|w|}$. More precisely, the first output $\hat{w}_0$ must belong to the language $(u_0 D_1^*) \otimes A^*$ and must synchronize the pair p_0 (*i.e.*, $[\![\hat{w}_0]\!] = p_0$). Such a word $\hat{w}_0$ exists because p_0 has length $\bar{y}_0 \in \langle \bar{x}_i, P_i \rangle \subseteq \pi(u_0 D_1^*)$. In a similar way, every subsequent output $\hat{w}_j$, for $j = 1, \ldots, |w|$, must be belong to the language

D_1^* and must synchronize the gap g_j (*i.e.*, $[\![\hat{w}_j]\!] = g_j$). Such a word $\hat{w}_j$ exists because g_j has length $|p_{j+1}| - |p_j|$, which belong to the linear set $\langle \bar{0}, P_i\rangle \subseteq \pi(D_1^*)$. By construction, the total output produced by $\mathcal{T}$ will be $\hat{w} = \hat{w}_0 \cdot \hat{w}_1 \cdots \hat{w}_{|w|}$, which belongs to $D \otimes \mathbb{A}^*$ and satisfies $[\![\hat{w}]\!] = p_{|w|} = [\![w]\!]$. ◀

▶ **Observation B.7.** *In general, $C \cap [D]_\pi$ and $C \setminus [D]_\pi$ are not regular if we do not impose any restriction. For example, $1^*2^* \cap (12)^* = \{1^n 2^n \mid n \in \mathbb{N}\}$.*

▶ **Corollary B.8.** *$C \cap [D^*u]_\pi \subseteq_{\text{REL}} D^*u$ for every regular C, D and word u.*

Proof. Let $C' = C \cap [D^*u]_\pi$. In order to be able to apply Proposition 11 for C' and D^*u, we need to prove that $\rho(\mathit{cycles}(C')) \subseteq \rho(\mathit{cycles}(D^*))$. After that, the result follows immediately from that proposition.

If $w \in \mathit{cycles}(C')$, then there exist words $\hat{u}, \hat{v}$ such that $\hat{u}w^*\hat{v} \subseteq C'$. Since $\pi(C') \subseteq \pi(D^*u)$, this implies that for all $n \in \mathbb{N}$, there exists $u_n \in D^*$ such that $\pi(\hat{u}w^n\hat{v}) = \pi(u_n u)$. Then it follows that $\inf \rho(\mathit{cycles}(D^*)) \leq \lim_{n\to\infty} \rho(u_n) = \lim_{n\to\infty} \rho(u_n u) = \lim_{n\to\infty} \rho(\hat{u}w^n\hat{v}) = \rho(w)$. In a similar way, one can prove that $\rho(w) \leq \sup \rho(\mathit{cycles}(D^*u))$. ◀

To complete the proof of Lemma 12, it only remains to provide the missing details for Claim 1:

▶ **Claim 1.** Every regular D^* is $=_{\text{REL}}$-equivalent to a finite union $\bigcup_i D_i^* u_i$, with finite D_i's.

Proof. Let $\langle \bar{x}_1, P_1\rangle \cup \cdots \cup \langle \bar{x}_n, P_n\rangle = \pi(D^*)$, and let D_i and u_i be so that $\pi(D_i^* u_i) = \langle \bar{x}_i, P_i\rangle$, for every $i = 1, \ldots, n$. We show that $D^* =_{\text{REL}} \bigcup_i D_i^* u_i$. The left-to-right containment holds by

$$\begin{aligned} D^* &= \bigcup_i (D^* \cap [D_i^* u_i]_\pi) && \text{(since } \pi(D^*) = \pi(\textstyle\bigcup_i D_i^* u_i)\text{)} \\ &\subseteq_{\text{REL}} \bigcup_i D_i^* u_i, && \text{(by Corollary B.8 plus P2)} \end{aligned}$$

and the right-to-left containment by Lemma 8 and the fact that for every i,

$$\begin{aligned} D_i^* u_i &= D_i^* u_i \cap [D^*]_\pi && \text{(since } \pi(D_i^* u_i) \subseteq \pi(D^*)\text{)} \\ &\subseteq_{\text{REL}} D^*. && \text{(by Corollary B.8)} \end{aligned}$$

◀

▶ **Lemma 13.** Every concat-star $C \subseteq \mathbb{2}^*$ of star-height 1 is $=_{\text{REL}}$-equivalence to a finite union $\bigcup_i C_i$ of simple languages.

Proof. First note that wlog we can assume that C is of the form

$$H_0 D_1^* H_1 D_2^* H_2 \cdots D_m^* H_m$$

where D_i^* are the heterogeneous components (non empty) and H_i are homogeneous concat-star languages (eventually empty). This is due to P5 and P2. Moreover, by using also P7, we can further assume that for all $i = 0, \ldots, m$ $H_i = 1^{k_i *} 2^{\hat{k}_i *}$ for some $k_i, \hat{k}_i$. We can

also assume that C is heterogeneous (otherwise the statement follows trivially), i.e. that $m \geq 1$. Since D_i^* are non empty heterogeneous components, for all $i = 1, \ldots, m$, there exist a heterogeneous word $w_i \in D_i^*$. Then

$$C = H_0 w_1^* D_1^* H_1 w_2^* D_2^* H_2 \cdots w_m^* D_m^* w_m^* H_m$$

and so, by what we said before and again P5 and P2, we can assume wlog that C is of the form $1^{k*}2^{\hat{k}*}w^*$ (the case $w^*1^{k*}2^{\hat{k}*}$ that we may need for the last homogeneous component is analogous). Then it only remains to prove the case where $C = 1^{k*}2^{\hat{k}*}w^*$ for $k, \hat{k} > 0$ and w a heterogeneous word. By P8, $1^{k*}2^{\hat{k}*} \subseteq_{\text{REL}} 1^{k*}(1^{k\hat{k}|w|_1}2^{\hat{k}k|w|_2})^*(2^{\hat{k}})^{<k|w|_2} \cup 2^{\hat{k}*}(1^{k\hat{k}|w|_1}2^{\hat{k}k|w|_2})^*(1^k)^{<\hat{k}|w|_1} \subseteq_{\text{REL}} 1^{k*}w^*(2^{\hat{k}})^{<k|w|_2} \cup 2^{\hat{k}*}w^*(1^k)^{<\hat{k}|w|_1}$. Then, it follows easily that $C = 1^{k*}2^{\hat{k}*}w^* =_{\text{REL}} 1^{k*}w^*(2^{\hat{k}})^{<k|w|_2} \cup 2^{\hat{k}*}w^*(1^k)^{<\hat{k}|w|_1}$ which concludes the proof. ◀

C Missing proofs to Section 6

▶ **Lemma 14.** For C a simple language and $D = \bigcup_i D_i$ finite union of simple languages, if $C \subseteq_{\text{Rel}} D$, then $C \xrightarrow{s.m.} D_i$ for some i. In particular, for C, D simple languages, if $C \subseteq_{\text{Rel}} D$, then $C \xrightarrow{s.m.} D$.

Proof. If C is homogeneous the statement follows immediately. Let us then assume that C is smooth heterogeneous of star-height 1.

The rough idea is to construct a relation $R \in \text{Rel}(C)$ in such a way that from $C \subseteq_{\text{Rel}} D$ one derives $R \in \text{Rel}(D)$, and from this, using suitable pumping arguments, one extracts a synchronizing morphism from C to some D_i. The relation R will depend on both languages C and D, but the underlying alphabet $\mathbb{A}$ will only depend on C. In fact, this latter dependency is mainly for simplifying the proof. Towards the end, we will explain how to avoid the dependency on C and construct a relation R over a fixed alphabet. This basically shows that semantics of $\subseteq_{\text{Rel}}$ with a fixed alphabet has the same characterization.

Consider a smooth heterogeneous language $C = C_1^* u_1 \cdots C_k^* u_k =_{\text{Rel}} C_1^* \cdots C_k^* u$, for $u = u_1 \cdots u_k$, and recall that there are no consecutive homogeneous components C_i^*, C_{i+1}^*. Without any loss of generality, assume that k is odd and C_j^* is heterogeneous for all even indices i.

We let D be a finite union of concat-star languages $D_i = D_{i,1}^* u_{i,1} \cdots D_{i,k_i}^* u_{i,k_i}$, and denote by $\mathcal{A}$ some automaton with h states that recognizes D.

The relation R is defined by taking into account the structure of the concat-star language $C_1^* \cdots C_k^* u$ and the number h of states of $\mathcal{A}$. Formally, we introduce the alphabet $\mathbb{A} = \{\tilde{a}_{i,j}, a_{i,j}, \tilde{b}_{i,j}, b_{i,j} \mid 1 \le i \le k, 1 \le j \le h\} \cup \{a_{k+1}\}$ and, for each $1 \le j \le k$, we denote by u_j^-, u_j^+ some words in C_j that have minimum and maximum Parikh ratios, respectively (minimum and maximum exists because C has star height 1). If j is odd, then we assume $u_j^- = u_j^+$. We also let $u_j^\sim \in C_j^*$ be any heterogeneous word for each even index j. We then define $R = [\![L]\!]$, where

$$L = L_{1,1} \cdots L_{1,h} \cdots L_{k,1} \cdots L_{k,h} \cdot (u \otimes a_{k+1}^{|u|}) \qquad \text{and}$$

$$L_{i,j} = \begin{cases} \left(u_j^- \otimes a_{i,j}^{|u_j^-|}\right)^* & \text{if } j \text{ is odd} \\ \left(u_j^\sim \otimes \tilde{a}_{i,j}^{|u_j^\sim|}\right)^* \cdot \left(u_j^- \otimes a_{i,j}^{|u_j^-|}\right)^* \cdot \left(u_j^\sim \otimes \tilde{b}_{i,j}^{|u_j^\sim|}\right)^* . \left(u_j^+ \otimes b_{i,j}^{|u_j^+|}\right)^* & \text{if } j \text{ is even.} \end{cases}$$

We exploit the fact that $R \in \text{Rel}(C')$ and $C' \subseteq_{\text{Rel}} D$ to derive $R \in \text{Rel}(D)$, and so $R = [\![L']\!]$ for some regular language $L' \subseteq D \otimes \mathbb{A}^*$. Let $\mathcal{B}$ be an automaton recognizing L' and, without loss of generality, assume that $\mathcal{B}$ is a refinement of $\mathcal{A}$ (this will be used later to transfer properties of cycles of $\mathcal{B}$ to properties of cycles of $\mathcal{A}$).

Now, consider the following word from the language L:

$$w = w_{1,1} \cdots w_{1,h} \cdots w_{k,1} \cdots w_{k,h} \cdot (u \otimes a_{k+1}^{|u|}) \qquad \text{with}$$

$$w_{i,j} = \begin{cases} \left(u_j^\sim \otimes \tilde{a}_{i,j}^{|u_j^\sim|}\right)^n \cdot \left(u_j^- \otimes a_{i,j}^{|u_j^-|}\right)^n \cdot \left(u_j^\sim \otimes \tilde{b}_{i,j}^{|u_j^\sim|}\right)^n . \left(u_j^+ \otimes b_{i,j}^{|u_j^+|}\right)^n & \text{if } j \text{ is even,} \\ \left(u_j^- \otimes a_{i,j}^{|u_j^-|}\right)^n & \text{if } j \text{ is odd,} \end{cases}$$

where n is chosen large enough so as to exceed the number of states of $\mathcal{B}$. Since $[\![L]\!] = R = [\![L']\!]$, we know that $\mathcal{B}$ accepts some word $\tilde{w}$ with $[\![\tilde{w}]\!] = [\![w]\!] = (w_1, w_2)$. Let γ be a successful run of $\mathcal{B}$ on $\tilde{w}$.

Now fix an arbitrary coordinate $\ell \in \mathbb{2}$, an index $1 \leq j \leq k$, and a letter $\tilde{c} \in \tilde{\mathbb{A}}_j = \{\tilde{a}_{i,j}, \tilde{b}_{i,j} \mid 1 \leq i \leq h\}$. Consider the first and last positions i_1, i_2 on γ that read (ℓ, c), and consider any arbitrary cycle between those positions i_1 and i_2. We claim that the cycle

1. reads only letters from $\mathbb{2} \times \{\tilde{c}\}$,
2. has Parikh ratio $\rho(u_j^{\sim})$.

Indeed, the positions of w_ℓ that carry the letter $\tilde{c}$ form a *factor* of w_ℓ. This implies that all the positions on γ between i_1 and i_2, and in particular, the positions of the cycle, either read $(\ell, \tilde{c})$ or read a symbol c' on the opposite coordinate $3-\ell$. If the cycle reads at least two different letters on coordinate $3-\ell$, then by pumping the cycle we obtain an inconsistency. Otherwise, if the cycle reads always the same letter c' on coordinate $3-\ell$, and $c' \neq \tilde{c}$, then by removing the cycle we would remove a positive number of occurrences of $(\ell, \tilde{c})$ but no occurrence of $(3-\ell, \tilde{c})$, thus obtaining a word $\tilde{w}' \in L'$ so that $\frac{\pi(\tilde{w}')(1,\tilde{c})}{\pi(\tilde{w}')(1,\tilde{c})+\pi(\tilde{w}')(2,\tilde{c})} \neq \rho(u_j^{\sim})$, and hence $[\![\tilde{w}']\!] \notin R$, which is a contradiction. So the cycle must read the same letter $\tilde{c}$ on coordinate $3-\ell$. Finally, if the Parikh ratio of the cycle is not precisely $\rho(u_j^{\sim})$, then by removing the cycle and by arguing as before, we would obtain a word $\tilde{w}' \in L'$ so that $[\![\tilde{w}']\!] \notin R$ (again a contradiction).

Now consider again an arbitrary coordinate $\ell \in \mathbb{2}$ and the maximal factors of w_ℓ that remain sandwiched between the previous factors identified by the symbols from $\bigcup_j \tilde{\mathbb{A}}_j$. By construction every such sandwiched factor reads the same letter c everywhere, for some $c \in \mathbb{A}_j = \{a_{i,j}, b_{i,j} \mid 1 \leq i \leq h\}$ and some $\leq j \leq k$. Consider the first and last positions i_1, i_2 of γ that read (ℓ, c). We claim that every cycle on γ between i_1 and i_2

1. reads only letters from $\mathbb{2} \times \{c\}$;
2. if $c = a_{i,j}$ for some $1 \leq i \leq h$, then it has Parikh ratio $\rho(u_j^-)$; if in addition $\rho(u_j^-) = 0$ (and hence $\ell = 2$), then it reads only letters $(2, a_{i,j})$; symmetrically, if $\rho(u_j^-) = 1$ (and hence $\ell = 1$), then it reads only letters $(1, a_{i,j})$;
3. if $c = b_{i,j}$ for some $1 \leq i \leq h$, then it has Parikh ratio $\rho(u_j^+)$; if in addition $\rho(u_j^+) = 0$ (and hence $\ell = 2$), then it reads only letters $(2, b_{i,j})$; symmetrically, if $\rho(u_j^+) = 1$ (and hence $\ell = 1$), then it reads only letters $(1, b_{i,j})$.

Note that if we are in the second case, *i.e.*$a_{i,j}$, but $\rho(u_j^-) \neq 0$, then, we could reach a contradiction by reasoning as we did before, that is, by removing the cycle and obtaining a word $\tilde{w}' \in L'$ so that $[\![\tilde{w}']\!] \notin R$. Similarly, it cannot happen that $c = b_{i,j}$ and $\rho(u_j^+) \neq 0$. So the interesting case here is when c is, for example, of the form $a_{i,j}$ and $\rho(u_j^-) = 0$. If on coordinate $3-\ell$ the cycle reads at least two different letters, we get an absurd as before. Otherwise, if it reads at least one letter $(3-\ell, c')$ for some $\tilde{c} \in \bigcup_j \tilde{\mathbb{A}}_j$, we get an absurd by removing the cycle and changing the ratio of letters $\tilde{c}$. Finally, if it reads at least one letter $(3-\ell, c')$ for some other $c' \in \mathbb{A}_{j'}$ with $j \neq j'$, it means that there is some $\tilde{c} \in \bigcup_j \tilde{\mathbb{A}}_j$ so that all the $(1, \tilde{c})$ letters occur before position i_1 and all the $(2, \tilde{c})$ letters occur after position i_2 on γ, or viceversa, the $(1, \tilde{c})$'s occur after i_2 and the $(2, \tilde{c})$'s occur before i_1. This is in contradiction with the previous claim on cycles inside factors of letters from $\bigcup_j \tilde{\mathbb{A}}_j$. The remaining cases are similar.

The previous claims imply that the letters

$$a_{1,1}, \ldots, a_{1,h}, \tilde{a}_{2,1}, a_{2,1}, \tilde{b}_{2,1}, b_{2,1} \ldots \tilde{a}_{2,h}, a_{2,h}, \tilde{b}_{2,h}, b_{2,h}, \ldots$$
$$\ldots, a_{k-1,1}, \ldots, a_{k-1,h}, \tilde{a}_{k,1}, a_{k,1}, \tilde{b}_{k,1}, b_{k,1} \ldots \tilde{a}_{k,h}, a_{k,h}, \tilde{b}_{k,h}, b_{k,h}$$

(assuming k is even), induce cycles in γ of Parikh ratios

$$\underbrace{\rho(u_1^-),\dots,\rho(u_1^-)}_{h \text{ times}},\underbrace{\rho(u_2^\sim),\rho(u_2^-),\rho(u_2^\sim),\rho(u_2^+),\dots,\rho(u_2^\sim),\rho(u_2^-),\rho(u_2^\sim),\rho(u_2^+)}_{4h \text{ times}},\dots$$
$$\dots,\underbrace{\rho(u_{k-1}^-),\dots,\rho(u_{k-1}^-)}_{h \text{ times}},\underbrace{\rho(u_k^\sim),\rho(u_k^-),\rho(u_k^\sim),\rho(u_k^+),\dots,\rho(u_k^\sim),\rho(u_k^-),\rho(u_k^\sim),\rho(u_k^+)}_{4h \text{ times}}$$

in this precise order. By choice of h, this means that there are k SCC's in $\mathcal{A}$, say $Q_1,\dots,Q_k$, so that

- Q_{i+1} is reachable from Q_i for every i, and
- Q_i contains a cycle with ratio $\rho(u_i^-)$ if i is odd, and cycles with ratios $\rho(u_i^-),\rho(u_i^-)$ if i is even.

From this it easy to conclude that there exists an index i and a synchronizing morphism from C to D_i.

It remains to explain how to modify the proof so as to have a relation R over a fixed alphabet. The main modification consists in removing the subscripts j from all the letters $\tilde{a}_{i,j}, a_{i,j}, \tilde{b}_{i,j}, b_{i,j}, a_{k+1}$ in $\mathbb{A}$, thus defining a new alphabet $\mathbb{A}' = \{\tilde{a}, a, \tilde{b}, b\}$ (by same principle, using more clumsy notation, we could even restrict to a binary alphabet). Accordingly, the langauge L is redefined as

$$L' = L'_{1,1}\cdots L'_{1,h}\cdots L'_{k,1}\cdots L'_{k,h}\cdot(u\otimes a^{|u|}) \qquad \text{where}$$

$$L'_{i,j} = \begin{cases} \left(u_j^- \otimes a^{|u_j^-|}\right)^* & \text{if } j \text{ is odd} \\ \left(u_j^\sim \otimes \tilde{a}^{|u_j^\sim|}\right)^* \cdot \left(u_j^- \otimes a^{|u_j^-|}\right)^* \cdot \left(u_j^\sim \otimes \tilde{b}^{|u_j^\sim|}\right)^* . \left(u_j^+ \otimes b^{|u_j^+|}\right)^* & \text{if } j \text{ is even.} \end{cases}$$

As before, we derive the existence of a regular language $L'' \subseteq D \otimes \mathbb{A}^*$ so that $[\![L']\!] = [\![L'']\!]$.

However, due to the modification, we cannot rely on the indexed letters to apply the remaining arguments. Nonetheless, we can overcome the problem by introducing a new automaton $\mathcal{C}$ that is a refinement of $\mathcal{B}$ (and hence also a refinement of $\mathcal{A}$) and that recognizes the old language L with the correctly indexed letters. The idea is that $\mathcal{C}$ reads indexed letters, while simulating $\mathcal{B}$ on the letters devoid of the indices, and checking at the same time the series of indices is correct. For this, $\mathcal{C}$ need to stores the current control state of the automaton $\mathcal{B}$ and the last letters c_1, c_2 from the parsed prefix of the input that were associated with each coordinate in $\mathbb{2}$. For example, when reading a letter $(1, a_{i,j})$, $\mathcal{C}$ simulates a transition of $\mathcal{B}$ on $(1, a)$, and checks that the last letter c_1 associated with coordinate 1 is either $a_{i,j}$ or $\tilde{a}_{i,j}$ (namely, the only two possible letters that could have preceded the current letter $(1, \tilde{a}_{i,j})$ in an arbitrary word from the languauge L).

From there, one follows basically the same arguments as before: first, construct a word $w \in L$ inducing cycles in a successful run of $\mathcal{C}$ (here n needs to exceed the number of states of $\mathcal{C}$). Then, argue that the cycles of $\mathcal{C}$ have the appropriate Parikh rations in the appropriate order. Finally, one transfers the latter properties to the successful runs of $\mathcal{B}$ and $\mathcal{A}$, so as to witness a synchronizing morphism. ◀

▶ **Lemma 16.** For every $p, q > 0$, finite $C \subseteq \mathbb{2}^*$, and $u_-, u_+ \in C$ so that $\rho(u_-) = \min \rho(C)$ and $\rho(u_+) = \max \rho(C)$, there exists a finite $C' \subseteq C^*$ so that

$$C^* =_{\text{REL}} (u_-^p \cup u_+^q)^* \cdot C'.$$

Before turning to the proof of Lemma 16, we need to first prove a few technical lemmas.

▶ **Lemma C.1.** *For every non-empty words $u_1, u_2, u_3 \in \mathbb{2}^*$ such that $\rho(u_1) \leq \rho(u_2) \leq \rho(u_3)$, there exist $k_1, k_2, k_3 \in \mathbb{N}$ such that $k_1 > 0$ and $\pi(u_2^{k_1}) = \pi(u_1^{k_2} \cdot u_3^{k_3})$. Moreover, if $\rho(u_1) < \rho(u_2)$, we can choose $k_3 > 0$ and if $\rho(u_2) < \rho(u_3)$, we can choose $k_2 > 0$.*

Proof. Linear algebra. ◀

▶ **Lemma C.2.** *For every finite language $C = \{u_-, u_+, u_1, \ldots, u_n\}$ with $\rho(u_-) = \min(\rho(C))$ and $\rho(u_+) = \max(\rho(C))$, there are $k_1, \ldots, k_n > 0$ so that*

$$C^* =_{\text{Rel}} (u_- \cup u_+)^* \cdot u_1^{<k_1} \cdots u_n^{<k_n}.$$

Proof. We proceed by induction on n. The base case $n = 0$ is immediate. For the inductive case, suppose $n > 0$, and let $k > 0$ and $k', k'' \geq 0$ be so that

$$\pi(u_n^k) = \pi(u_-^{k'} \cdot u_+^{k''}) \tag{1}$$

by Lemma C.1. Thus,

$$\begin{aligned}
C^* &= ((\hat{C}^* \cdot u_n)^k)^* \cdot (\hat{C}^* \cdot u_n)^{<k} \cdot \hat{C}^* && \text{for } \hat{C} = C \setminus \{u_n\} \\
&=_{\text{Rel}} \hat{C}^* \cdot (\hat{C}^* \cdot u_n)^{<k} \cdot \hat{C}^* && \text{(by (1) } \textit{cum} \text{ P5, P6, P2, P3)} \\
&=_{\text{Rel}} \hat{C}^* \cdot \hat{C}^* \cdot u_n^{<k} \cdot \hat{C}^* =_{\text{Rel}} \hat{C}^* \cdot u_n^{<k} && \text{(by P5)} \\
&=_{\text{Rel}} (u_- \cup u_+)^* \cdot u_1^{<k_1} \cdots u_{n-1}^{<k_{n-1}} \cdot u_n^{<k}, && \text{(by inductive hypothesis on } \hat{C})
\end{aligned}$$

proving the statement. ◀

▶ **Lemma C.3.** *For every $p > 0$, $u_1, u_2 \in \mathbb{2}^*$ we have*

$$(u_1 \cup u_2)^* =_{\text{Rel}} (u_1^p \cup u_2)^* \cdot u_1^{<p}.$$

Proof. We have

$$\begin{aligned}
(u_1 \cup u_2)^* &= u_2^*(u_1 u_2^*)^{p*}(u_1 u_2^*)^{<p} \\
&=_{\text{Rel}} (u_1^p \cup u_2)^*(u_1 u_2^*)^{<p} =_{\text{Rel}} (u_1^p \cup u_2)^* u_2^* u_1^{<p} && \text{(by P5)} \\
&= (u_1^p \cup u_2)^* u_1^{<p} && \blacktriangleleft
\end{aligned}$$

▶ **Corollary C.4** (of Lemma C.3)**.** *For every $p, q > 0$, $u_1, u_2 \in \mathbb{2}^*$ we have*

$$(u_1 \cup u_2)^* =_{\text{Rel}} (u_1^p \cup u_2^q)^* \cdot u_2^{<q} \cdot u_1^{<p}.$$

Proof. It follows easily from Lemma C.3 (applied twice) plus P2. ◀

Proof of Lemma lem:newnormalform. Straightforward application of Lemma C.2 and Corollary C.4 plus P2. ◀

▶ **Proposition 7.** For all simple $C, D \subseteq \mathbb{2}^*$, $C \subseteq_{\text{Rel}} D$ iff $\pi(C) \subseteq \pi(D)$ and $C \xrightarrow{s.m.} D$.

Proof. The left-to-right direction follows from P6 and Lemma 14. For the right-to-left direction, if C is homogeneous, the fact that $\pi(C) \subseteq \pi(D)$ yields $C \subseteq_{\text{Rel}} D$ by P8. Suppose then that C is heterogeneous. For any concat-star $C_1^* u_1 \cdots C_n^* u_n$, note that $C_1^* u_1 \cdots C_n^* u_n =_{\text{Rel}} C_1^* \cdots C_n^* \cdot u$ for $u = u_1 \cdots u_n$ and $\pi(C_1^* u_1 \cdots C_n^* u_n) = \pi(C_1^* \cdots C_n^* \cdot u)$. Thus, we can assume wlog that C and D are of the form $C = C_1^* \cdots C_n^* u$ and $D = D_1^* \cdots D_m^* v$, and let

$u = u_1 \cdots u_n$, $v = v_1 \cdots v_m$. Note also that, by a similar argument, we can assume wlog that $C_i \neq \{\varepsilon\}$ for all $i = 1, \ldots, n$.

Since all C_i are finite, one can take $w_{i,-}, w_{i,+}$ be the minimum and maximum Parikh-ratio words of each C_i respectively —note that they could have the same Parikh-ratio, or even be the same word.

Let us fix some i. Since $\rho(C_i^*) \subseteq \rho(D_{j_i}^*)$ for some j_i, there are words $\hat{w}_1, \hat{w}_2 \in D_{j_i}^*$ so that $\rho(w_{i,r}) = \rho(\hat{w}_r)$ for $r \in \{-, +\}$ or, in other words, there are $p, q > 0$ so that $\pi(w_{i,-}^p) = \pi(\hat{w}_1)$ and $\pi(w_{i,+}^q) = \pi(\hat{w}_2)$.

By Lemma 16, we have

$$\begin{aligned} C_i^* &=_{\text{REL}} (w_{i,-}^p \cup w_{i,+}^q)^* \cdot C_i' && \text{(for a finite } C_i' \subseteq C_i^*) \\ &\subseteq_{\text{REL}} D_{j_i}^* \cdot C_i' \end{aligned}$$

Thus,

$$\begin{aligned} C_1^* \cdots C_n^* u &\subseteq_{\text{REL}} D_{j_1}^* \cdot C_1' \cdots D_{j_n}^* \cdot C_n' \cdot u \\ &=_{\text{REL}} D_{j_1}^* \cdots D_{j_n}^* \cdot C_1' \cdots C_n' \cdot u \\ &\subseteq_{\text{REL}} D_1^* \cdots D_m^* C_1' \cdots C_n' \cdot u && \text{(by monotonicity of s.m.)} \\ &\subseteq_{\text{REL}} D_1^* \cdots D_m^* v \end{aligned}$$

proving the statement. For the last step, note that $\pi(c_1 \cdots c_n u) \in \pi(D_1^* \cdots D_n^* v)$ for all $c_i \in C_i'$ and so the containment follows from P6 and P2. ◀