Inferring Unions of the Pattern Languages By The Most Fitting Covers

Yen Kaow Ng [a,1] Takeshi Shinohara [b]

[a] *Kyushu Institute of Technology, Graduate School of Computer Science and Systems, Iizuka, 820, Japan*

[b] *Kyushu Institute of Technology, Department of Artificial Intelligence, Iizuka, 820, Japan*

Abstract

We are interested in learning unions of the pattern languages in the limit from positive data using strategies that guarantee some form of minimality during the learning process. It is known that for any class of languages with so-called finite elasticity, any learning strategy that finds a language that is minimal with respect to inclusion (MINL) ensures identification in the limit. We consider a learning strategy via another form of minimality — the minimality in the number of elements that are shorter than a specified length. A search for languages with this minimality is possible in many cases, and the search can be adapted to identify any class where every language in the class has a characteristic set within the class. We compare solutions using this strategy to those from MINL to illustrate how we may obtain solutions that fulfill both notions of minimality. Finally, we show how the results are relevant using subclasses of the pattern languages.

Key words: Unions of languages, Pattern Languages, Inductive Inference, Most Fitting Covers

1 Introduction

In this paper we study some strategies for learning from positive data [6] the unions of the pattern languages [1] and its subclasses. We are interested in getting hypotheses that not only converge to the presented languages in

Email addresses: `kalngyk@daisy.ai.kyutech.ac.jp` (Yen Kaow Ng), `shino@ai.kyutech.ac.jp` (Takeshi Shinohara).

[1] Supported by the Japanese Government Scholarship of the Ministry of Education, Science, Sports, Culture and Technology of Japan.

the limit, but also guarantee other desirable properties, such as some form of minimality.

One such minimality is that with respect to set inclusion. Given a class of languages where the membership is uniformly decidable (called an indexed family of recursive languages [1]), and a property known as *finite elasticity* [21] is fulfilled, every language in the class can be identified in the limit by finding, for any sample received at each stage, a language which (1) includes all the samples, and (2) is minimal with respect to set inclusion among the other languages in the class that contain all the samples. An language that fulfills (1) and (2) is known as MINL in the literature [1,18].

The unions of up to a bounded number of the non-erasing pattern languages is particularly interesting in this case because the class has finite elasticity. That is, the ability to compute MINL is sufficient for identifying languages in the class [4]. Arimura *et. al.* devised an efficient algorithm (MMG) that computes MINL for the bounded unions of languages where a condition known as *compactness with respect to containment* can be fulfilled [4]. Sato *et. al.* showed that for the *regular patterns* (that is, patterns where each variable appears at most once), the property can be fulfilled when the number of pattern languages allowed in a union is less than a bound that depends on the alphabet size [17]. In general, the MINL problem is difficult to compute, partly due to the difficulty in deciding inclusion between infinite languages [9,10].

On the other hand, it may be desirable to achieve another form of minimality that is not guaranteed by MINL. We illustrate this with an example using the regular pattern languages. Let $S = \{\text{“}cabcc\text{”}, \text{“}cbcac\text{”}\}$, and consider the language generated by the pattern “$*a*$” (that is, the set of strings obtained by replacing the variables “$*$”s with any non-empty string over a, b, and c). Since no other language that includes S is a proper subset of it, “$*a*$” is minimal with respect to inclusion and hence is a potential output of MINL. However, another possible solution of MINL, “$*bc*$” generates less number of strings of any given length than “$*a*$”, and may be preferred in some applications [5,14].

Based on this we propose another approach to minimality. We propose to find, among the hypotheses that are consistent with the given sample, for one that contains the least number of elements of up to a specified length. More precisely, an instance of the problem would ask for a set of languages by specifying the following: (1) a class $\mathcal{L}$ of languages from which the output is drawn, (2) k, the number of languages allowed in the output, (3) a set S of samples that the output languages must together contain, and (4) a length ℓ, where the output must contain the least number of elements of up to length ℓ, among all the sets of languages that fulfill (1)–(3). We denote an instance of the problem specified in such a way $\mathbf{FC}(\mathcal{L}, S, k, \ell)$. (The efficient computability of such a problem for some subclasses of the pattern languages has been studied in [15].)

Now let us look at how we may use an algorithm that computes $\mathbf{FC}(\mathcal{L}, S, k, \ell)$ for learning. First of all we want to base our conjectures solely on the output from a computation of $\mathbf{FC}(\mathcal{L}, S, k, \ell)$, without using knowledge obtainable via other means (e.g. information possibly encoded in the indexing of the languages). In this sense, all the solutions in $\mathbf{FC}(\mathcal{L}, S, k, \ell)$ are treated as equals. We then consider if it is possible to bound ℓ to some value. We can, perhaps, set ℓ to the length of the longest element in S. This however creates a problem: Suppose there are two languages L, $L' \in \mathcal{L}$ where L is finite, $L \subseteq L'$, and the two languages contains exactly the same elements of within the length (m say) of the longest element in L. In this case, given a complete representation of the elements in L, both L and L' will be in $\mathbf{FC}(\mathcal{L}, L, k, m)$, and hence the output of $\mathbf{FC}(\mathcal{L}, L, k, m)$ alone is insufficient for us to decide on the correct conjecture L. This problem, however, does not arise for a class that consists only of infinite languages, since there is no bound to the length of the elements of any language. Of course, the problem is also avoided if we simply allow ℓ to be extended indefinitely. Hence, in this paper we consider the following two cases: (Case 1) allow only infinite languages in the class, and compute $\mathbf{FC}(\mathcal{L}, S, k, m)$ where m is the length of the longest element in S; and (Case 2) allow both finite and infinite languages in the class, but allow ℓ to extend indefinitely beyond the lengths of the samples.

We found similar results in both cases: such a learner learns a class of languages if and only if, for every language L in the class, there exists a finite set $S \subseteq L$ that in some way, witnesses the minimality of L in the FC sense. While this may seem exceedingly restrictive, it turns out that any class of languages where each language in it has a *characteristic set within the class* [3] fulfills this requirement. This includes any class with finite elasticity [16,11], e.g. the class of *bounded unions of non-erasing pattern languages* [21]. Also, even though the full class of the erasing pattern languages does not have finite elasticity, many of its interesting subclasses do, e.g. the *erasing regular pattern languages* [18].

We then turn to the question of obtaining unions of the pattern languages that are at the same time minimal with respect to MINL and FC, since both notions of minimality are frequently desirable properties to have. We observe that for any S, if there is a way to decide a length n for which every union of up to k languages in $\{L \in \mathcal{L} \mid S \cap L \neq \emptyset\}$ has a characteristic set consisting of elements of no longer than n, then every solution in $\mathbf{FC}(\mathcal{L}, S, k, n)$ also fulfills the minimality of MINL. This allows us to use previous results by others [17,20] to show that for the (erasing and non-erasing) regular pattern languages, if the number k of unions allowed is less than a bound dictated by the alphabet size, then for any $\mathcal{L}$ and S, any output from $\mathbf{FC}(\mathcal{L}, S, k, \ell)$ also fulfills the minimality required by MINL, if ℓ is as long as some factor dependent on the length of the elements in S.

Through the same method we also show how solutions that fulfill both MINL and FC can be obtained for unions of the (erasing and non-erasing) languages generated respectively by the *substring patterns* [19], the *subsequence patterns* [7], and the patterns where only one variable occurrence is allowed in each pattern.

2 Preliminaries

The symbol N denotes the set of natural numbers. The symbol N^+ denotes the set of positive natural numbers. Cardinality of a set S is written $card(S)$. A *word* over a non-empty alphabet A is a finite string of symbols taken from A. The *empty word* is a null string and denoted by ϵ. A^*, A^+, denote the sets of all the words and non-empty words over A. and words of length n or less over A, respectively. For a set S, $S^{\leq n}$ denotes the words of length up to n in S.

Let Σ be a finite alphabet with $card(\Sigma) \geq 2$, and $V = \{x_1, x_2, \ldots\}$ be a countable alphabet disjoint from Σ. Elements in Σ are called *constants* and elements in V are called *variables*. A non-empty word over $V \cup \Sigma$ is called a *pattern*, and the length of a pattern p is denoted $|p|$. Each constant or variable has unit length. For a set of patterns P, the length of the longest word in it is denoted $\lceil P \rceil$ while the length of the shortest word in it is denoted $\lfloor P \rfloor$.

A *substitution* is a homomorphism from patterns to patterns that maps every constant to itself. The image of a pattern p under a substitution θ is written $p\theta$. A substitution is *erasing* if it maps some or all variables to the empty word, and it is *non-erasing* otherwise. Unless stated otherwise, substitutions are non-erasing in this paper. We write $p \preceq q$ iff $p = q\theta$ for some substitution θ, and we write $p \preceq_\epsilon q$ iff $p = q\theta$ for some erasing substitution θ. We write $p \prec q$ just in case $p \preceq q$ but not $q \preceq p$. Given two sets of patterns P and Q, we write $P \sqsubseteq Q$ iff for each $p \in P$, $p \preceq q$ for some $q \in Q$. $P \sqsubset Q$ just in case $P \sqsubseteq Q$ but not $Q \sqsubseteq P$. Similarly we define the relations $\prec_\epsilon$, $\sqsubseteq_\epsilon$ and $\sqsubset_\epsilon$ from $\preceq_\epsilon$. To improve readability we will often abbreviate the erasing version of some notations and results when they can be straightforwardly derived from the non-erasing case.

A *regular pattern* is a pattern where each variable appears at most once [18]. A pattern is a *substring pattern* just in case it is of the form "$x_1 s x_2$" and $s \in \Sigma^+$. A pattern is a *subsequence pattern* just in case it is of the form "$x_1 a_1 x_2 a_2 \ldots x_n a_n x_{n+1}$" where the subscripted "a"s are elements of Σ [7]. We denote the sets of all regular patterns, substring patterns and subsequence patterns by **RP**, **Psub**, and **Psubseq**, respectively.

A *language* is a set of words over Σ. The *pattern language* of a pattern p, written $L(p)$, is the set $\{w \in \Sigma^* \mid w \preceq p\}$. For a set P of patterns, $L(P)$ denotes the language $\bigcup_{p \in P} L(p)$ and $\mathcal{L}(P)$ denotes the class of languages $\{L(p) \mid p \in P\}$. The symbols L_ϵ and $\mathcal{L}_\epsilon$ can be defined similarly.

2.1 Inductive inference

We present the concepts in inductive inference [6] in this section. A *sequence* is an ordered presentation of elements in Σ^* where repetitions are allowed. For each sequence σ, $content(\sigma)$ denotes the set of elements that appeared in σ. SEQ denotes the set of all finite sequences. A *text* of a language L is an infinite sequence $w_1, w_2, \ldots$ such that $\{w_i \mid i \in N\} = L$. The initial segment of length n of a text T is denoted $T[n]$. An *inductive inference machine* (**IIM**) is an algorithmic device which computes a mapping from SEQ into N. We say that an **IIM** M *converges* on a text T to i just in case there exists $n \in N^+$ such that for all $n' \geq n$, M outputs i on input $T[n']$.

A collection of non-empty languages $\mathcal{L} = \{L_i \mid i \in N\}$ is *an indexed family of recursive languages* just in case there exists a computable function f such that for each $i \in N$ and for each $w \in \Sigma^*$, $f(i, w) = 1$ if $w \in L_i$ and $f(i, w) = 0$ otherwise. A family $L_1, L_2, \ldots$ is said to be ***TxtEx**-identified by an **IIM** M* just in case for each L_i, and for each text T for L_i, M converges on T to j, and $L_j = L_i$. $\mathcal{L}$ is *learnable* just in case some **IIM** M **TxtEx**-identifies it. A *finite tell-tale* for a language L within a class $\mathcal{L}$ is a finite set $S \subseteq L$ such that there exists no language $L' \in \mathcal{L}$ where $S \subseteq L' \subset L$. An indexed family of recursive languages $\{L_i \mid i \in N\}$ is learnable just in case there exists a procedure, effective in each i, that enumerates a finite tell-tale for L_i within the class [2].

A collection of languages $\mathcal{L}$ has *finite thickness* just in case for each $w \in \Sigma^*$, $\{L \in \mathcal{L} \mid w \in L\}$ is finite. A collection of languages $\mathcal{L}$ has *infinite elasticity* just in case there exists an infinite sequence of pairwise distinct words $w_0, w_1, w_2, \ldots$ from Σ^* and an infinite sequence of pairwise distinct languages $A_1, A_2, \ldots$ such that for each $k \in N^+$, $\{w_i \mid i < k\} \subseteq A_k$, but $w_k \notin A_k$. $\mathcal{L}$ is said to have *finite elasticity* just in case $\mathcal{L}$ does not have infinite elasticity [21,13]. A *characteristic set for a language L within a class $\mathcal{L}$* is a finite set S_L such that for each $L' \in \mathcal{L}$, $S_L \subseteq L' \Rightarrow L \subseteq L'$ [3]. Sato [16] and Kobayashi *et. al.* [11] independently showed that a class of languages $\mathcal{L}$ has finite elasticity if and only if every language $L \subseteq \Sigma^*$ has a characteristic set within $\mathcal{L}$. [2]

3 The fittest cover problem and its use as a strategy for inductive inference

We now formalize our notion of minimality.

Definition 1 *Given a class $\mathcal{L}$ of languages, k, $\ell \in N^+$, and a finite non-empty set of examples $S \subseteq \bigcup_{L\in\mathcal{L}} L$, the* Fittest Cover Problem *(written as* $\mathbf{FC}(\mathcal{L}, S, k, \ell)$*) is to find, among every set $\mathcal{L}'$ of up to k languages in* $\{L \in \mathcal{L} \mid$

[2] Note that for a class with finite elasticity, *every* language (that is, including languages outside of $\mathcal{L}$) has a characteristic set within $\mathcal{L}$.

$L \cap S \neq \emptyset\}$ *where* $S \subseteq \bigcup_{L\in\mathcal{L}'} L$*, for one that minimizes* $card(\bigcup_{L\in\mathcal{L}'} L^{\leq \ell})$.

As an example consider the sample $S = \{$"*cabcc*", "*cbcac*"$\}$. It is easy to verify that $L($"x_1bcx_2"$)$ is a solution for $\mathbf{FC}(\mathcal{L}(\mathbf{RP}), S, 1, \lceil S \rceil)$, but $L($"x_1ax_2"$)$ is not. Note that both are solutions for MINL.

For a given class of patterns $\mathcal{P}$, we say that a set of patterns $P \subseteq \mathcal{P}$ is a solution for $\mathbf{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)$ just in case $L(P)$ is a solution for $\mathbf{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)$ That is, we identify a pattern set P with the language it generates whenever the context is clear.

Definition 2 *Given a class* $\mathcal{P}$ *of patterns,* $k, \ell \in N^+$ *and* $S \subseteq \Sigma^*$*,* $\mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell) = \{P \subseteq \mathcal{P} \mid P$ *is a solution for* $\boldsymbol{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)\}$*, and* $\mathcal{FC}(\mathcal{L}_\epsilon(\mathcal{P}), S, k, \ell) = \{P \subseteq \mathcal{P} \mid P$ *is a solution for* $\boldsymbol{FC}(\mathcal{L}_\epsilon(\mathcal{P}), S, k, \ell)\}$.

We want to consider the use of FC as a strategy in inductive inference, that is, we let a learner conjecture one of the solutions obtained in an instance of the problem. An immediate concern regarding the use of such a strategy is that its minimality is partial. That is, while the minimality in MINL is with respect to Σ^*, an FC solution is only minimal for up to a specified length. In this section we show how the requirement of such a minimality restricts learning. To simplify the discussion in this section we use only $\mathbf{FC}(\mathcal{L}, S, 1, \ell)$, that is, the case without union. This does not make the discussion any less relevant to the study of unions since the finite unions of an indexed family is also an indexed family.

Given an indexed family of recursive languages $\mathcal{L} = \{L_i \mid i \in N\}$, and $\lambda \in N$, we let ***InferByFC***$(\ell = \lambda)$ be a computation that, upon given each initial segment $T[n]$ of a text T, conjectures the language with the smallest index among the solutions for $\mathbf{FC}(\mathcal{L}, content(T[n]), 1, \lambda)$.

In this paper we study the two cases: (1) ***InferByFC***$(\ell = \lceil S \rceil)$, which uses the solutions for $\mathbf{FC}(\mathcal{L}, content(T[n]), 1, \lceil content(T[n]) \rceil)$, and (2) ***InferByFC***$(\ell = n)$, which uses the solutions for $\mathbf{FC}(\mathcal{L}, content(T[n]), 1, n)$. We refer to the learner of (1) by ***FCLnrConLen***, and that of (2) by ***FCLnrTxtLen***.

Note that the least index is used in ***InferByFC***$(\ell = \lambda)$ to simplify convergence and its use does not add a dependency on the numberings for the cases we consider (where ℓ extends indefinitely), as the results will show. In the case that $\mathbf{FC}(\mathcal{L}, content(T[n]), 1, \lambda)$ is not computable due to that $\mathcal{L}$ is infinite, we may alternatively compute for $\mathbf{FC}(\{L_0, L_1, \ldots, L_n\}, content(T[n]), 1, \lambda)$. Since the correct language is included in the hypothesis space in the limit, this does not weaken the strength of the learner in terms of **TxtEx**-identification.

Definition 3 *Given a class* $\mathcal{L}$ *of languages and a language* L*, we say that a finite set* $S \subseteq L$ *is a*
(1) witness of minimality for L within $\mathcal{L}$ *just in case there exists* $\ell \in N$ *such that for all* $L' \in \mathcal{L}$*,* $S \subseteq L' \Rightarrow$ *for all* $\ell' \geq \ell$*,* $card(L^{\leq \ell'}) \leq card(L'^{\leq \ell'})$*,*
(2) witness of neighborhood minimality for L within $\mathcal{L}$ *just in case for every* S' *where* $S \subseteq S' \subseteq L$ *and for every* $L' \in \mathcal{L}$*,* $S' \subseteq L' \Rightarrow card(L^{\leq \lceil S' \rceil}) \leq card(L'^{\leq \lceil S' \rceil})$.

As mentioned in the introduction, ***FCLnrConLen*** does not identify a class of languages where there exists L_1 and L_2 such that $L_2^{\leq\lceil L_1\rceil} = L_1$, hence we consider classes that consists only of infinite languages in the following characterization of ***FCLnrConLen***.

Theorem 4 *__FCLnrConLen__ __TxtEx__-identifies an indexed family of infinite recursive languages $\mathcal{L} = \{L_i \mid i \in N\}$ if and only if for each i there exists a witness of neighborhood minimality for L_i within $\mathcal{L}$.*

Proof. ($\Leftarrow$) For $L_i \in \mathcal{L}$, let T be any text for L_i, and let S_i be a witness of neighborhood minimality for L within $\mathcal{L}$. Let $\ell \in N$ and finite $A_i \subseteq \Sigma^*$ be such that for each language $L_j \neq L_i$ where $j < i$, (1) if $L_i \subset L_j$, then $card(L_i^{\leq \ell}) < card(L_j^{\leq \ell})$, (2) otherwise A_i contains an element in $L_i - L_j$. Let x_i be an element longer than ℓ in L_i. Let $n \in N$ (where we furthermore require that $n \geq i$ if ***FCLnrConLen*** is made to compute **FC** from the class $\{L_0, \ldots, L_n\}$ instead of $\mathcal{L}$) be such that $A_i \cup S_i \cup \{x_i\} \subseteq content(T[n])$. (Note that we do not require such n to be recursively determined.) Then, for all $T[n']$ where $n' \geq n$, in the case that $j < i$, either (1) $L_j = L_i$, (2) $content(T[n]) \not\subseteq L_j$, or (3) $card(L_i^{\leq n}) < card(L_j^{\leq n})$; and in the case that $j > i$, $card(L_i^{\leq n}) \leq card(L_j^{\leq n})$. From this it is not difficult to see that ***FCLnrConLen*** will converge to some $i' \leq i$ where $L_{i'} = L_i$ after reading $T[n]$.

($\Rightarrow$) Given a language L_i where no such witness of neighborhood minimality exists, we show how to construct a text T for L_i for which ***FCLnrConLen*** fails to converge on to an index for a language for L_i. At each stage l of the construction we look for a finite set $S^l \subseteq L_i$ and sequence σ^l where every S^{l+1} includes $L_i^{\leq\lceil S^l\rceil}$ and every sequence σ^{l+1} begins with σ^l. At stage 0, let σ^0 be the empty sequence and let $S^0 = \emptyset$. At every stage $l+1$, we find a language $L_j \neq L_i$ and S^{l+1} where $L_i^{\leq\lceil S^l\rceil} \subseteq S^{l+1} \subseteq L_j \cap L_i$ and $card(L_i^{\leq\lceil S^{l+1}\rceil})$ $> card(L_j^{\leq\lceil S^{l+1}\rceil})$. This search will always succeed, because otherwise $L_i^{\leq\lceil S^l\rceil}$ is a witness of neighborhood minimality for L_i within $\mathcal{L}$. Let σ^{l+1} be a sequence (of length $\geq j$ if ***FCLnrConLen*** is made to compute **FC** from the class $\{L_0, \ldots, L_n\}$ instead of $\mathcal{L}$) extending σ^l where $content(\sigma^{l+1}) = S^{l+1}$. Let T be a text where each initial portion begins with a sequence σ^l for some l. Since at every stage $l+1$, σ^{l+1} contains all the elements in L_i of up to length $\lceil S^l\rceil$, T is a text for L_i. However, ***FCLnrConLen*** fails to converge on T to an index for L_i. ∎

Note that the existence of such witnesses of neighborhood minimality suffices for their effective enumeration in the limit.

Corollary 5 *There exists a learnable indexed family $\mathcal{L}$ of infinite recursive languages not __TxtEx__-identified by __FCLnrConLen__.*

Proof. Let $a \in \Sigma$ and let a^n denote the word in $\{a\}^*$ of length n. For each i, $j \in N^+$, $j \geq 2$, let $L_{ij} = \{a^n \mid n \leq i \ \vee\ n \equiv i \pmod j\}$. Let $\mathcal{L} = \{L_{ij} \mid i, j \in N^+, j \geq 2\}$. For each i, $j \in N$, let S_{ij} be a finite set including $\{a^n \mid n \leq i\}$ and $\{a^{i+j}\}$. It is easy to verify that S_{ij} is a finite tell-tale for L_{ij} and hence $\mathcal{L}$ is learnable. We now show that there is no witness of neighborhood minimality for any language in the class. For any i, j, let S be any finite subset of L_{ij}. Let i', $j' \in N$ be such that: (1) $i' \geq \max\{i, \lceil S \rceil\}$, (2) $i' + j' \equiv i \pmod{j}$, and (3) $i + (i' + j' - i)/j > i' + 1$. Let $S' = S \cup \{a^{i'+j'}\}$. By (1), $S' \subseteq L_{i'j'}$; and by (2), $S' \subseteq L_{ij}$. However, there are at least $i + (i' + j' - i)/j$ ($> i' + 1$) elements in $L_{ij}^{\leq i'+j'}$, but only $i' + 1$ elements in $L_{i'j'}^{\leq i'+j'}$. ∎

The following can be shown using essentially the same argument as that in proving Theorem 4. We skip the details.

Theorem 6 ***FCLnrTxtLen TxtEx****-identifies an indexed family of recursive languages $\mathcal{L} = \{L_i \mid i \in N\}$ if and only if for each i there exists a witness of minimality for L_i within $\mathcal{L}$.*

Corollary 7 *If each language in an indexed family $\mathcal{L}$ of recursive languages has a characteristic set within $\mathcal{L}$,* ***FCLnrTxtLen TxtEx****-identifies $\mathcal{L}$.*

Note that each language in the class of all finite languages has itself as a characteristic set, and hence the class can be learned via ***FCLnrTxtLen***. For an example of a class learnable by ***FCLnrTxtLen*** where no language in it has a characteristic set within the class, let $w_0, w_1, \ldots$ be an enumeration of Σ^* and consider the class $\mathcal{L} = \{\Sigma^* - \{w_i\} \mid i \in N\}$.

Corollary 8 *Given an indexed family $\mathcal{L}$ of recursive languages with finite elasticity and $k \in N$,* ***FCLnrTxtLen TxtEx****-identifies the class of unions of up to k languages in $\mathcal{L}$.*

The class of unbounded unions of the erasing languages of substring patterns has infinite elasticity, but each language in it has a characteristic set within the class [19].

4 Comparing FC to MINL and MMG

The problems MINL and MMG [4] have been introduced in the past as learning strategies. MINL, as mentioned, aims to find languages that are minimal with respect to set inclusion. MMG, on the other hand, looks for patterns that are minimal with respect to the '$\sqsubseteq$' relation. In this section we see how the FC problem relates to these problems, being motivated by the possibility of obtaining solutions that fulfill more than one of these notions of minimality. We first formally state the MINL and MMG problems. The following definition extends straightforwardly to the erasing pattern languages by replacing the symbols L with L_ϵ, and $\mathcal{L}$ with $\mathcal{L}_\epsilon$.

Definition 9 *Given a class of patterns $\mathcal{P}$, a finite non-empty set $S \subset \Sigma^*$ and $k \in N^+$,*
(1) The problem ***MINL***$(\mathcal{L}(\mathcal{P}), S, k)$ *is to find a set of at most k patterns $P \subseteq \mathcal{P}$ such that $S \subseteq L(P)$ and there are no other set of up to k patterns $P' \subseteq \mathcal{P}$ where $S \subseteq L(P')$ such that $L(P') \subset L(P)$.*
(2) The problem ***MMG***$(\mathcal{L}(\mathcal{P}), S, k)$ *is to find a set of at most k patterns*

$P \subseteq \mathcal{P}$ such that $S \subseteq L(P)$ and there are no other set of up to k patterns $P' \subseteq \mathcal{P}$ where $S \subseteq L(P')$ such that $P' \sqsubset P$.
(3) $\mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k) = \{P \subseteq \mathcal{P} \mid P$ is a solution for $\boldsymbol{MINL}(\mathcal{L}(\mathcal{P}), S, k)\}$.
(4) $\mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k) = \{P \subseteq \mathcal{P} \mid P$ is a solution for $\boldsymbol{MMG}(\mathcal{L}(\mathcal{P}), S, k)\}$.

Note that $\mathcal{FC}$ and $\mathcal{MINL}$ are defined to be classes of patterns rather than classes of languages solely to simplify the comparison with $\mathcal{MMG}$. The results in this section regarding $\mathcal{FC}$ and $\mathcal{MINL}$ can be easily adapted to remove the dependency on a pattern class $\mathcal{P}$.

For any class $\mathcal{P}$ of patterns, since (from [1]) $(\forall P,\ Q \subseteq \mathcal{P})\ [P \sqsubseteq Q \Rightarrow L(P) \subseteq L(Q)]$, a solution P for $\mathbf{MINL}(\mathcal{L}(\mathcal{P}), S, k)$ is either a solution for $\mathbf{MMG}(\mathcal{L}(\mathcal{P}),\ S, k)$, or there exists some solution Q for $\mathbf{MMG}(\mathcal{L}(\mathcal{P}),\ S, k)$ where $L(P) = L(Q)$ (i.e. $P \sqsubset Q \Rightarrow L(P) \subset L(Q) \vee L(P) = L(Q)$). On the other hand, $\mathcal{MMG}(\mathcal{L}(\mathcal{P}),\ S, k) \subseteq \mathcal{MINL}(\mathcal{L}(\mathcal{P}),\ S, k)$ if for all $P,\ Q \in \mathcal{P}$, $L(P) \subseteq L(Q) \Rightarrow P \sqsubseteq Q$, as noted in [4]. The following states this condition more precisely.

Proposition 10 *If for every sets of up to k patterns $P,\ Q \subseteq \mathcal{P}$ we have $L(P) \subseteq L(Q) \Rightarrow P \sqsubseteq Q$, then $\mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k) = \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$.*

Proof. We want to show $L(P) \subset L(Q) \Leftrightarrow P \sqsubset Q$. $(\Rightarrow)$ is given by $L(P) \subset L(Q)$ (by the hypothesis)$\Rightarrow P \sqsubseteq Q$, and $P \sqsubseteq Q \wedge Q \sqsubseteq P \Rightarrow L(P) = L(Q)$. We now show $P \sqsubset Q \Rightarrow L(P) \subset L(Q)$. Let P and Q be sets of patterns from $\mathcal{P}$ where $P \sqsubset Q$. We argue that $L(Q) - L(P) \neq \emptyset$. Suppose otherwise, then $P \sqsubset Q \Rightarrow L(P) \subseteq L(Q)$ (since $L(Q) - L(P) = \emptyset$)$\Rightarrow L(Q) = L(P)$ (by the hypothesis)$\Rightarrow Q \sqsubseteq P$, a contradiction. ∎

The following states necessary and sufficient conditions for the solutions of $\mathbf{FC}(\mathcal{L}, S, k, \ell)$ to be solutions for $\mathbf{MMG}(\mathcal{L}, S, k)$ and $\mathbf{MINL}(\mathcal{L}, S, k)$.

Proposition 11 *Given $k, \ell \in N^+$, class $\mathcal{P}$ of patterns and a finite non-empty $S \subseteq \Sigma^*$. Let $\mathcal{P}' = \{p \in \mathcal{P} \mid L(p) \cap S \neq \emptyset\}$. Then,*
***1**. $\mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell) \subseteq \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$ if and only if, for each solution $P \in \mathcal{FC}(\mathcal{L}(\mathcal{P}),\ S, k, \ell)$, every set of up to k patterns $Q \subseteq \mathcal{P}'$ where $S \subseteq L(Q)$ has $L(P) \subset L(Q) \Rightarrow L(Q)^{\leq \ell} - L(P)^{\leq \ell} \neq \emptyset$.*
***2**. $\mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell) \subseteq \mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k)$ if and only if, for each solution $P \in \mathcal{FC}(\mathcal{L}(\mathcal{P}),\ S, k, \ell)$, every set of up to k patterns $Q \subseteq \mathcal{P}'$ where $S \subseteq L(Q)$ has $P \sqsubset Q \Rightarrow L(Q)^{\leq \ell} - L(P)^{\leq \ell} \neq \emptyset$.*

Proof. Part **1**. $(\Rightarrow)$ Let $P \in \mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)$. Suppose there exists a set of up to k patterns $Q \subseteq \mathcal{P}'$ where $S \subseteq L(Q)$, $L(P) \subset L(Q)$, and $L(Q)^{\leq \ell} - L(P)^{\leq \ell} = \emptyset$, then clearly, $Q \in \mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)$ but $Q \notin \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$ (as witnessed by P). Then, $\mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell) \not\subseteq \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$.

$(\Leftarrow)$ Towards a contradiction let $Q \in \mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell) - \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$. We then have:

(A) $Q \subseteq \mathcal{P}'$ (since $Q \in \mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)$), and

(B) there exists $L(P) \subset L(Q)$ where $S \subseteq L(P)$, that is, P witnesses that $Q \notin \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$. We assume $P \subseteq \mathcal{P}'$, since otherwise we may remove the patterns in P that are not in $\mathcal{P}'$. Since P contains strictly less elements than Q, $P \in \mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)$.

Then, by the Proposition, $L(Q)^{\leq \ell} - L(P)^{\leq \ell} \neq \emptyset$, and hence $Q \notin \mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)$, a contradiction.

Part **2** is proven similarly. ∎

Propositions 10 and 11 extend staightforwardly, with the same arguments, to the erasing pattern languages. The following result relates the classes $\mathcal{FC}$ and $\mathcal{MINL}$ using the characteristic sets, and extends immediately to the erasing case.

Theorem 12 *Given a class $\mathcal{P}$ of patterns, a finite non-empty $S \subseteq \Sigma^*$ and $k \in N^+$. Let $\mathcal{P}' = \{p \in \mathcal{P} \mid L(p) \cap S \neq \emptyset\}$ and let $\mathcal{L}^k = \{L(P) \mid P \subseteq \mathcal{P}' \wedge card(P) \leq k\}$. If there exists $\ell \in N$ such that every language $L(P) \in \mathcal{L}^k$ has a characteristic set $S_P \subseteq L(P)^{\leq \ell}$ within $\mathcal{L}^k$, then $\mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell) \subseteq \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$.*

Proof. Let $P \in \mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \ell)$. Suppose there exists $L(Q) \in \mathcal{L}^k$ where $L(P) \subset L(Q)$ and $L(Q)^{\leq \ell} - L(P)^{\leq \ell} = \emptyset$. Let $S_Q \subseteq L(Q)^{\leq \ell}$ be a characteristic set of $L(Q)$ within $\mathcal{L}^k$. Since $L(Q)^{\leq \ell} - L(P)^{\leq \ell} = \emptyset$, we would have $S_Q \subseteq L(P)$ and hence $L(Q) \subseteq L(P)$, a contradiction. Hence no such $L(Q)$ exists, and therefore the Theorem holds by Part 1 of Proposition 11. ∎

Hence it suffices that we specify a sufficiently large ℓ, depending only on the maximum length of the elements in the characteristic sets, in the FC problem to obtain MINL solutions. A caveat here is that even though a characteristic set is known to exist for every non-erasing pattern language within the class (since the class has finite elasticity), such maximum length of the elements in a characteristic set cannot be recursively determined for every non-erasing pattern language. The reason is that, if such lengths can be decided, then inclusion for any two patterns can be decided — by recursively generating their constituent elements of up to that length. However, Jiang *et. al.* has shown that the problem of inclusion between any two given non-erasing pattern languages is undecidable [10].

Nevertheless, the same problem does not arise for the regular patterns nor its bounded unions. It has been shown that under specific conditions, such sufficient lengths for the elements of characteristic sets can be computed for these regular pattern languages. We present these results next.

5 Languages fulfilling both FC and MINL

In this section we apply Theorem 12 on some subclasses of the pattern languages to demonstrate how we may be able to find, from these subclasses, languages containing a set of sample that fulfill both FC and MINL notions of minimality.

5.1 Regular patterns

We first consider the languages generated by the regular patterns. We will first state the results from [17] and [20], and then present the corollaries relevant to the present work.

Definition 13 ([17,20]) *For each pattern p, we denote by*
(a) $S_n(p)$ the set of all the words obtainable from p by substituting each variable in p with a non-empty element from $(\Sigma^*)^{\leq n}$.
(b) $S_n^\epsilon(p)$ the set of all the words obtainable from p by substituting each variable in p with a string in Σ^* of length n.
(c) For a set of patterns P, $S_n(P) = \bigcup_{p \in P} S_n(p)$ and $S_n^\epsilon(P) = \bigcup_{p \in P} S_n^\epsilon(p)$.

Sato *et. al.* showed that it is possible to use S_1 and S_2 as characteristic sets for the bounded unions of the regular pattern languages, when the alphabet size is sufficiently larger than the number of unions allowed, as follows.

Theorem 14 (Theorem 11 and 20 in [17]) *Let $k \in N^+$. For every set of patterns $P \subseteq \boldsymbol{RP}$ and up to k patterns $Q \subseteq \boldsymbol{RP}$,*
(1) if $card(\Sigma) \geq 2k+1$, then the following three statements are equivalent:
(i) $S_1(P) \subseteq L(Q)$, (ii) $P \sqsubseteq Q$, (iii) $L(P) \subseteq L(Q)$.
(2) if $k \geq 3$ and $card(\Sigma) \geq 2k-1$, then the following three statements are equivalent:
(i) $S_2(P) \subseteq L(Q)$, (ii) $P \sqsubseteq Q$, (iii) $L(P) \subseteq L(Q)$.

The equivalence of (i) and (iii) in Theorem 14 shows that $S_1(P)$ and $S_2(P)$ are characteristic sets for any collection of languages from $\mathcal{L}(\mathbf{RP})$ within the class $\{L(P) \mid P \subseteq \mathbf{RP} \wedge card(P) \leq k\}$, under the respective conditions. The following Corollary follows.

Corollary 15 *For every finite non-empty $S \subseteq \Sigma^*$ and $k \in N^+$.*
(1) if $card(\Sigma) \geq 2k+1$, $\mathcal{FC}(\mathcal{L}(\boldsymbol{RP}), S, k, \lceil S \rceil) \subseteq \mathcal{MINL}(\mathcal{L}(\boldsymbol{RP}), S, k) = \mathcal{MMG}(\mathcal{L}(\boldsymbol{RP}), S, k)$.
(2) if $card(\Sigma) \geq 2k-1$ and $k \geq 3$, $\mathcal{FC}(\mathcal{L}(\boldsymbol{RP}), S, k, 2\lceil S \rceil) \subseteq \mathcal{MINL}(\mathcal{L}(\boldsymbol{RP}), S, k) = \mathcal{MMG}(\mathcal{L}(\boldsymbol{RP}), S, k)$.

Proof. (1) We first note that given any S, the set $\mathcal{P} = \{p \in \mathbf{RP} \mid L(p) \cap S \neq \emptyset\}$ consists only of patterns of length up to $\lceil S \rceil$. $S_1(P)$ where $P \subseteq \mathcal{P}$ hence contains only elements of up to length $\lceil S \rceil$. Theorem 12 and 14 then give us $\mathcal{FC}(\mathcal{L}(\mathbf{RP}), S, k, \lceil S \rceil) \subseteq \mathcal{MINL}(\mathcal{L}(\mathbf{RP}), S, k)$. That $\mathcal{MINL}(\mathcal{L}(\mathbf{RP}), S, k) = \mathcal{MMG}(\mathcal{L}(\mathbf{RP}), S, k)$ is from Proposition 10 and the equivalence of (ii) and (iii) in Theorem 14. Part (2) is shown similarly. The equivalence between $\mathcal{MINL}$ and $\mathcal{MMG}$ is due to the equivalence of (ii) and (iii) in Theorem 14. ∎

For the erasing regular pattern languages, Uemura and Sato [20] obtained the following.

Theorem 16 (Theorem 2 and 3 in [20]) *Let $k \in N^+$. For every collection of up to k patterns $P, Q \subseteq \boldsymbol{RP}$,*

(1) if $card(\Sigma) \geq 2k+1$*, then the following three statements are equivalent:*
(i) $S_1^\epsilon(P) \subseteq L_\epsilon(Q)$*, (ii)* $P \sqsubseteq_\epsilon Q$*, (iii)* $L_\epsilon(P) \subseteq L_\epsilon(Q)$*.*
(2) if $card(\Sigma) \geq k+2$*, then the following three statements are equivalent:*
(i) $S_2^\epsilon(P) \subseteq L_\epsilon(Q)$*, (ii)* $P \sqsubseteq_\epsilon Q$*, (iii)* $L_\epsilon(P) \subseteq L_\epsilon(Q)$*.*

Corollary 17 *For every finite non-empty* $S \subseteq \Sigma^*$ *and* $k \in N^+$*,*
(1) if $card(\Sigma) \geq 2k+1$*, then* $\mathcal{FC}(\mathcal{L}_\epsilon(\boldsymbol{RP}), S, k, 2\lceil S\rceil+1) \subseteq \mathcal{MINL}(\mathcal{L}_\epsilon(\boldsymbol{RP}), S, k)$*;*
(2) if $card(\Sigma) \geq k+2$*, then* $\mathcal{FC}(\mathcal{L}_\epsilon(\boldsymbol{RP}), S, k, 3\lceil S\rceil+2) \subseteq \mathcal{MINL}(\mathcal{L}_\epsilon(\boldsymbol{RP}), S, k)$ $= \mathcal{MMG}(\mathcal{L}_\epsilon(\boldsymbol{RP}), S, k)$*.*

Proof. Without loss of generality we consider only patterns where no variables appear consecutively, because replacing any such consecutive variables with a single variable does not alter the corresponding erasing pattern language. Denote this set of patterns by $\mathbf{RP}'$. Now given any S, the set $\mathcal{P} = \{p \in \mathbf{RP}' \mid L(p) \cap S \neq \emptyset\}$ consists only of patterns of length up to $2\lceil S\rceil + 1$, each containing at most $\lceil S\rceil + 1$ variables. Hence for any $P \subseteq \mathcal{P}$, strings in $S_1^\epsilon(P)$ are of length at most $2\lceil S\rceil + 1$, while strings in $S_2^\epsilon(P)$ are of length at most $3\lceil S\rceil + 2$. Theorem 12 and 16 then give us the subset relations between $\mathcal{FC}$ and $\mathcal{MINL}$. The equivalence between $\mathcal{MINL}$ and $\mathcal{MMG}$ is due to the equivalence of (ii) and (iii) in Theorem 16. ∎

5.2 Substring patterns and subsequence patterns

For some subclasses of the regular patterns, stronger results relating FC to MINL and MMG can be obtained.

Lemma 18 *Suppose* $card(\Sigma) \geq 2$ *and* $\mathcal{P} \in \{\boldsymbol{Psub}, \boldsymbol{Psubseq}\}$*. For any* $p \in \mathcal{P}$*,*
(a) there exists a word $w \in L(p)$ *of length* $\lfloor L(p)\rfloor$ *such that for all* $q \in \mathcal{P}$ *where* $p \not\preceq q$*,* w *is not in* $L(q)$*.*
(b) there exists a word $w \in L_\epsilon(p)$ *of length* $\lfloor L_\epsilon(p)\rfloor$ *such that for all* $q \in \mathcal{P}$ *where* $p \not\preceq_\epsilon q$*,* w *is not in* $L_\epsilon(q)$*.*

Proof. We first show Part (a) for **Psub**. Let $p = x_1ux_2$ where x_1, $x_2 \in V$ and $u \in \Sigma^+$. We claim that any word $w = aub$ for some (possibly equal) a, $b \in \Sigma$ is not in $L(q)$ if $p \not\preceq q$. Towards a contradiction let $q = y_1vy_2$ where y_1, $y_2 \in V$, $v \in \Sigma^+$, and suppose $w \in L(q)$. Since $w \in L(q)$, $w = aw_1vw_2b$ for some w_1, $w_2 \in \Sigma^*$. However, the substitution θ where $y_1\theta = x_1w_1$ and $y_2\theta = w_2x_2$ then witnesses that $p \preceq q$, a contradiction. Part (b) for **Psub** can be shown similarly by letting $w = u$.

We now show Part (a) for **Psubseq**. Let $p = x_1a_1x_2a_2 \ldots a_nx_{n+1}$ where $x_1, \ldots, x_{n+1} \in V$ and $a_1, \ldots, a_n \in \Sigma$. Let $w = p\theta$ where θ is a substitution transforming each x_i to a_i for $1 \leq i \leq n$, and transforming x_{n+1} to a_n. Now for any $q \in$ **Psubseq**, if $w \preceq q$, it is not difficult to see that q must be of the form $y_1a_{i_1}y_2a_{i_2} \ldots a_{i_k}y_{k+1}$ where $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ and $y_1, \ldots, y_{k+1}$ $\in V$. However, then $p \preceq q$. Part (b) for **Psubseq** can be shown with $w = p\theta'$ where θ' is a substitution which erases all the variables. ∎

Note that Lemma 18 implies $L(P) \subseteq L(Q)$ or $L_\epsilon(P) \subseteq L_\epsilon(Q) \Rightarrow P \sqsubseteq Q$. In the following Lemma, for a set of patterns P, $c(P)$ denotes the maximum of $\{\lfloor L(p) \rfloor \mid p \in P\}$.

Lemma 19 *Suppose card*$(\Sigma) \geq 2$ *and* $\mathcal{P} \in \{\boldsymbol{Psub}, \boldsymbol{Psubseq}\}$. *For any two sets of finite* $P, Q \subseteq \mathcal{P}$,
(a) $P \sqsubset Q \Rightarrow L(Q)^{\leq c(Q)} - L(P)^{\leq c(Q)} \neq \emptyset$,
(b) $L(P) \subset L(Q) \Rightarrow L(Q)^{\leq c(Q)} - L(P)^{\leq c(Q)} \neq \emptyset$, *and*
(c) $P \sqsubset Q \Leftrightarrow L(P) \subset L(Q)$.

Proof. We first proof Part (a). Suppose there exists $P, Q \subseteq \mathcal{P}$ where $P \sqsubset Q$ and $L(Q)^{\leq c(Q)} - L(P)^{\leq c(Q)} = \emptyset$. Since $L(Q)^{\leq c(Q)} - L(P)^{\leq c(Q)} = \emptyset$, by Lemma 18, $q \in Q$ implies $q \preceq p$ for some $p \in P$, and hence $Q \sqsubseteq P$, a contradiction. For Part (c), that $P \sqsubset Q \Rightarrow L(P) \subset L(Q)$ is from Part (a) and the fact that $P \sqsubset Q \Rightarrow L(P) \subseteq L(Q)$. We now show $L(P) \subset L(Q) \Rightarrow P \sqsubset Q$. By Lemma 18, $L(P) \subset L(Q)$ implies that $P \sqsubseteq Q$. If furthermore $Q \sqsubseteq P$ then $L(P) = L(Q)$, contradicting $L(P) \subset L(Q)$. Hence we have $P \sqsubset Q$. Part (b) follows from (a) and (c). ∎

Note also that $L(P) \subseteq L(Q) \Rightarrow P \sqsubseteq Q$ is sufficient to show Lemma 19(c). Due to the identity between the non-erasing and erasing case in Lemma 18, Lemma 19 extends immediately to the erasing case, with L replaced by L_ϵ, $\mathcal{L}$ replaced by $\mathcal{L}_\epsilon$, and $c(P)$ redefined as $\{\lfloor L_\epsilon(p) \rfloor \mid p \in P\}$.

Theorem 20 *Suppose card*$(\Sigma) \geq 2$. *Given* $\mathcal{P} \in \{\boldsymbol{Psub}, \boldsymbol{Psubseq}\}$, *for every finite non-empty* $S \subseteq \Sigma^*$ *and* $k \in N^+$,
(1) $\mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \lceil S \rceil) \subseteq \mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k) = \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$;
(2) $\mathcal{FC}(\mathcal{L}_\epsilon(\mathcal{P}), S, k, \lceil S \rceil) \subseteq \mathcal{MMG}(\mathcal{L}_\epsilon(\mathcal{P}), S, k) = \mathcal{MINL}(\mathcal{L}_\epsilon(\mathcal{P}), S, k)$.

Proof. We first note that given any S, every pattern in $\mathcal{P}' = \{p \in \mathcal{P} \mid L(p) \cap S \neq \emptyset\}$ must have their shortest element not longer than $\lceil S \rceil$. Hence for any $P, Q \subseteq \mathcal{P}'$, $P \sqsubset Q$ (by Lemma 19(a))$\Rightarrow L(Q)^{\leq c(Q)} - L(P)^{\leq c(Q)} \neq \emptyset$ $\Rightarrow L(Q)^{\leq \lceil S \rceil} - L(P)^{\leq \lceil S \rceil} \neq \emptyset$. Then by Proposition 11, $\mathcal{FC}(\mathcal{L}(\mathcal{P}), S, k, \lceil S \rceil) \subseteq \mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k)$. That $\mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k) = \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S, k)$ is from Lemma 19(c).

The erasing case is shown identically. ∎

Note that Theorem 20 continues to hold in the case of unbounded unions.

In general, $\mathcal{MINL}(\mathcal{L}, S, k) - \mathcal{FC}(\mathcal{L}, S, k, \lceil S \rceil)$ is non-empty. Examples such as the following can be easily obtained.

Example 21 *Let* $\Sigma = \{a, b\}$. *(For* **Psub***) Let* $P = \{$"x_1aaax_2", "x_1bbbx_2"$\}$, $Q = \{$"x_1abax_2", "x_1bbax_2"$\}$, *and* $S = L(P) \cap L(Q) \cap (\Sigma^*)^{\leq 9}$. *It can be verified that both* $P, Q \in \mathcal{MINL}(\mathcal{L}(\boldsymbol{Psub}), S, 2)$, *but* $P \notin \mathcal{FC}(\mathcal{L}(\boldsymbol{Psub}), S, 2, \lceil S \rceil)$. *(For* **Psubseq***) Let* $P = \{$"$x_1ax_2ax_3ax_4$", "$x_1bx_2bx_3bx_4$"$\}$, $Q = \{$"$x_1ax_2bx_3ax_4$", "$x_1bx_2bx_3ax_4$"$\}$, *and* $S = L(P) \cap L(Q) \cap (\Sigma^*)^{\leq 8}$. *It can be verified that both* $P, Q \in \mathcal{MINL}(\mathcal{L}(\boldsymbol{Psubseq}), S, 2)$, *but* $P \notin \mathcal{FC}(\mathcal{L}(\boldsymbol{Psubseq}), S, 2, \lceil S \rceil)$.

Nevertheless, on some very specific samples, for **Psub** and **Psubseq**, the FC problem is no different from the MINL or MMG problem, as the following result shows. (Note that the following holds similarly for the erasing case.)

Proposition 22 *Suppose* $card(\Sigma) \geq 2$ *and* $\mathcal{P} \in \{\boldsymbol{Psub}, \boldsymbol{Psubseq}\}$. *For each finite* $P \subseteq \mathcal{P}$, *there exists a finite non-empty set* $S_P \subseteq \Sigma^*$ *such that for all* $k \geq card(P)$ *and all finite* S' *where* $S_P \subseteq S' \subseteq L(P)$, $\mathcal{FC}(\mathcal{L}(\mathcal{P}), S', k) = \mathcal{MINL}(\mathcal{L}(\mathcal{P}), S', k) = \mathcal{MMG}(\mathcal{L}(\mathcal{P}), S', k)$.

Proof. Let θ be an arbitrary substitution that transforms each variable to a letter in Σ. Given $P \subseteq \mathcal{P}$, we first note from the proof of Lemma 18 that for each $p \in P$, $p\theta \in L(Q)$ implies that $p \preceq q$ for some $q \in Q$. Hence, let $S_P = \{p\theta \mid p \in P\}$. Now for any $Q \subseteq \mathcal{P}$, $S_P \subseteq L(Q)$ implies that for each $p \in P$, there exists some $q \in Q$ such that $p \preceq q$. Hence for any superset S of S_P, $S \subseteq L(Q) \Rightarrow P \sqsubseteq Q$. Now if $P \sqsubset Q$, then $Q \notin \mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k)$. Hence $\mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k) = \{P' \mid P' \sqsubseteq P \wedge P \sqsubseteq P'\}$. That each $P' \in \mathcal{MMG}(\mathcal{L}(\mathcal{P}), S, k)$ generates the same language then gives us the result through Theorem 20. ∎

5.3 Patterns with at most 1 variable occurrence

In the subclasses discussed so far, any FC or MMG solution is at the same time a solution for MINL. We now show a class of pattern languages where an FC solution is also a solution for MINL, but an MMG solution is not necessarily so.

Let $\mathbf{P}_{\leq 1}$ be the class of patterns where only one variable occurrence is allowed in each pattern. It is easy to see that the condition $(\forall P, Q \in \mathbf{P}_{\leq 1})[L(P) \subseteq L(Q) \Rightarrow P \sqsubseteq Q]$ does not hold when $k \geq card(\Sigma)$. As an example, let $\Sigma = \{a, b\}$, then $L(\text{“}x_1aa\text{”}) \subset L(\text{“}ax_1a\text{”}) \cup L(\text{“}bx_1a\text{”})$, but neither “$x_1aa$” $\preceq$ “ax_1a” nor “x_1aa” $\preceq$ “bx_1a” holds. Hence a solution of MMG is not necessarily a solution for MINL. In the following, for any rational number r, $\lfloor r \rfloor$ denotes the largest natural number smaller than r.

Theorem 23 *For any finite* $P \subseteq \boldsymbol{P}_{\leq 1}$, $S_t(P)$ *where* $t = 2\lfloor \frac{k-1}{card(\Sigma)-1} \rfloor + 2$ *is a characteristic set for* $L(P)$ *within* $\{L(Q) \mid Q \subseteq \boldsymbol{P}_{\leq 1} \wedge card(Q) \leq k\}$.

Proof. Let variables $V = \{x\}$ and constants $\Sigma = \{a_1, \ldots, a_n\}$. We define two sets of substitutions,

(1) $\Theta_1 = \{\theta \mid x\theta = a_i x a_j, 1 \leq i, j \leq n\}$, and

(2) $\Theta_2 = \{\theta \mid x\theta \in \Sigma^{\leq 2}\}$.

For any set of patterns P and for any set of substitutions Θ, we let $\Theta \cdot P$ denote the set of patterns $\{p\theta \mid p \in P \wedge \theta \in \Theta\}$. Furthermore, for $l \in N$, we let $(\Theta)^l \cdot P$ denote the set of patterns obtainable from l applications of substitutions from Θ on patterns in P, that is, for example, $(\Theta)^2 = \Theta \cdot (\Theta \cdot P)$. Note that $S_{2l}(P) = \bigcup_{0 \leq i \leq l-1} \Theta_2 \cdot ((\Theta_1)^i \cdot P)$.

Claim 24 *For any* $P \subseteq \boldsymbol{P}_{\leq 1}$, *and finite* $Q \subseteq \boldsymbol{P}_{\leq 1}$, $L(P) \subseteq L(Q)$ *if and only if there exists* $l \in N$ *such that (I)* $(\Theta_1)^l \cdot P \sqsubseteq Q$, *and (II)* $\bigcup_{0 \leq i \leq l-1} \Theta_2 \cdot ((\Theta_1)^i \cdot P) \subseteq L(Q)$.

Proof of Claim. ($\Leftarrow$) is not difficult to see. We show ($\Rightarrow$). We claim that (I) must hold for some $l \in N$. Let $Q = \{u_i x v_i \mid 1 \leq i \leq card(Q) \wedge u_i, v_i \in \Sigma^*\}$ be, and let m be the longest length of such u_i and v_i. Now consider $P' = (\Theta_1)^{m+1} \cdot P$. Clearly, every pattern in P' is of the form uxv where $u, v \in \Sigma^*$, $|u| > m$ and $|v| > m$. Suppose $P' \not\sqsubseteq Q$, then there exists $uxv \in P'$, u, $v \in \Sigma^*$ such that for every $u_i x v_i \in Q$, u does not begin with u_i or v does not end in v_i, and hence $u a_1 v \notin L(Q) \Rightarrow L(P) \not\subseteq L(Q)$, a contradiction. Now note that if (II) fails for any $l \in N$ then $L(P) \not\subseteq L(Q)$ also. Claim follows from these observations. □

The following computation uses the principle in the Claim to decide, given $p \in \mathbf{P}_{\leq 1}$ and finite $Q \subseteq \mathbf{P}_{\leq 1}$, if $L(p) \subseteq L(Q)$. At each stage $l \in N$, the computation is as follows:

stage 0:

  If $p \preceq q$ for some $q \in Q$, output true.
  Otherwise let $P_1 = \{p\}$ and $Q_1 = Q$.
  Enter stage 1.

stage l:

  If $\Theta_2 \cdot P_l \not\subseteq L(Q_l)$, output false.
  If $\Theta_1 \cdot P_l \sqsubseteq Q_l$, output true.
  Otherwise let
    $P_{l+1} \subseteq \Theta_1 \cdot P_l$ be the set of patterns not refinable from any q in Q_l,
    $Q_{l+1} \subseteq Q_l$ be the set of patterns $q \in Q$ where either
      1. q contains a variable, but refines to nothing in $\Theta_1 \cdot P_l$, or
      2. q contains no variable, and $|q| \geq |p| + 2l$.
  Enter stage $l + 1$.

Note that:

(A) At any stage l, patterns in $Q_l - Q_{l+1}$ do not refine to any pattern in $\Theta_1 \cdot P_{l'}$ and $\Theta_2 \cdot P_{l'}$ for any $l' > l$.

Hence they are excluded from evaluation at subsequent stages. From the computation, we also see that at any stage l, if $Q_l = \emptyset$, then either

(1) $P_l = \emptyset$, in which case $(\Theta_1)^{l-1} \cdot \{p\} \sqsubseteq Q$ and (by Claim 24) $L(p) \subseteq L(Q)$, or

(2) $\Theta_2 \cdot P_l \not\subseteq L(Q)$, hence $S_{2l}(p) \not\subseteq L(Q)$.

Hence we have:

(B) $Q_l = \emptyset \Rightarrow [S_{2l}(p) \subseteq L(Q) \Leftrightarrow L(p) \subseteq L(Q)]$.

We will show that for any $l > \lfloor (k-1)/(n-1) \rfloor$, either $Q_l = \emptyset$ or $S_{2l}(p) \not\subseteq L(Q)$, which together with (B) completes the proof.

For each stage l, let $Q'_l = Q_l - Q_{l+1}$, that is, Q'_l are the patterns removed from Q_l at stage l. We assume each $q \in Q$ which contains a variable refine to some $p' \in \Theta_1 \cdot P_m$ for some $m \in N$, since otherwise $L(q) \cap L(p) = \emptyset$, in which case q can be discarded from consideration. That is,

(C) Each $q \in Q$, with or without a variable occurrence, is in Q'_m for some $m \in N$.

We also note the following:

(D) $card(Q) \leq k$.

(E) $card(\Theta_1 \cdot P_l) = n^2 card(P_l)$.

(F) Each $q \in Q'_m$ refines to at most n patterns in $\Theta_1 \cdot P_m$, and at most $n+1$ patterns in $\Theta_2 \cdot P_m$.

By (C)–(F), we can see that there is an upper bound to the number of stages that can be reached for any set Q before $card(Q_l) \leq n$. Let Q be such that this number of stages reached is maximized, and let ℓ be the stage when $card(Q_\ell) \leq n$. We note that such ℓ is maximized when $card(Q) = k$ and $card(P_m)$ is minimized for each $m < \ell$. Hence for each m, $card(P_m) = 1$ and $card(\Theta \cdot P_m) = n^2$. By (F), this gives us that $card(\Theta_1 \cdot P_m) = n^2 \leq n \cdot card(Q'_m) + 1 \Rightarrow card(Q'_m) \geq n - 1$. Hence at least $(\ell-1)(n-1)$ patterns from Q are in some Q'_m in the stages $m < \ell$, and so there are $k-(\ell-1)(n-1)$ patterns in Q_ℓ. We consider the following two cases:

(Case I) $card(Q_\ell) = n$. Note that $k - (\ell-1)(n-1) = n \Rightarrow \ell = (k-1)/(n-1)$. In this case $Q_{\ell+1} = \emptyset$.

(Case II) $card(Q_\ell) < n$. Note that $k-(\ell-1)(n-1) = m$ for some $m < n$ $\Rightarrow \ell = (k+n-1-m)/(n-1) \Rightarrow (k-1)/(n-1) \leq \ell \leq (k-1)/(n-1)+1$. Since $card(\Theta_2 \cdot P_\ell) = n^2 + n$, by Condition (F), this case has $\Theta_2 \cdot P_\ell \not\subseteq L(Q_\ell)$.

Hence as claimed, for $l > \lfloor (k-1)/(n-1) \rfloor$, either $Q_l = \emptyset$ or $S_{2l}(p) \not\subseteq L(Q)$. In the case that $Q_l = \emptyset$, as argued, $S_{2l}(p) \subseteq L(Q) \Leftrightarrow L(p) \subseteq L(Q)$. Hence $S_{2l}(p) \subseteq L(Q) \Leftrightarrow L(p) \subseteq L(Q)$ holds in general when $l > \lfloor (k-1)/(n-1) \rfloor$, and we are done. The case for any set of patterns $P \subseteq \mathbf{P}_{\leq 1}$ follows immediately. ∎

The erasing case has an identical result.

Theorem 25 *For any finite $P \subseteq \boldsymbol{P}_{\leq 1}$, $S_t(P)$ where $t = 2\lfloor \frac{k-1}{card(\Sigma)-1} \rfloor + 2$ is a characteristic set for $L_\epsilon(P)$ within $\{L_\epsilon(Q) \mid Q \subseteq \boldsymbol{P}_{\leq 1} \wedge card(Q) \leq k\}$.*

Proof. The proof is the same as Theorem 23, and it can be verified that each of the conditions remains valid, with the following minor differences.

1. For the computation at stage l, in constructing Q_{l+1}, the condition for the patterns without variables is now $|q| \geq |p| + 2l - 1$.

2. Condition (F) becomes: *Each $q \in Q'_m$ refines to at most n patterns in $\Theta_1 \cdot P_m$, and at most $n+2$ patterns in $\Theta_2 \cdot P_m$.* This does not change the result in (Case II). (Note that $card(\Theta_2 \cdot P_m) = n^2 + n + 1$ for the present case.) ∎

Noting that Theorem 12 requires only the characteristic sets to be within a subset of the target class which intersect non-emptily with the samples, we obtain the following result.

Theorem 26 *Given any finite non-empty set $S \subseteq \Sigma^*$. Let $\boldsymbol{Q}_{\leq 1} = \{p \in \boldsymbol{P}_{\leq 1} \mid L(p) \cap S \neq \emptyset\}$. For any finite $P \subseteq \boldsymbol{Q}_{\leq 1}$,*
(1) $S_t(P)$ where $t = 2\min\{\lceil S \rceil, \lfloor \frac{k-1}{card(\Sigma)-1} \rfloor\} + 2$ is a characteristic set for $L(P)$

within $\{L(Q) \mid Q \subseteq \boldsymbol{Q}_{\leq 1} \wedge card(Q) \leq k\}$.
(2) $S_t(P)$ *where* $t = 2\min\{\lceil S \rceil + 1, \lfloor \frac{k-1}{card(\Sigma)-1} \rfloor\} + 2$ *is a characteristic set for* $L_\epsilon(P)$ *within* $\{L_\epsilon(Q) \mid Q \subseteq \boldsymbol{Q}_{\leq 1} \wedge card(Q) \leq k\}$.

Proof. We show Part (1) first. We follow the same proof in Theorem 23. Recall that each $q \in Q$ is in Q'_m for some $m \in N$, otherwise $L(q) \cap L(p) = \emptyset$. We furthermore note that, $q \in Q'_m \Rightarrow |q| \geq m$. That is, any pattern q of length up to n which is not in some Q'_m where $m \leq n$ has $L(q) \cap L(p) = \emptyset$. Since every pattern in $\mathbf{Q}_{\leq 1}$ is of length at most $\lceil S \rceil$, each $q \in \mathbf{Q}_{\leq 1}$ which intersects non-emptily with $L(p)$ must be in some Q'_m where $m \leq \lceil S \rceil$. Discarding the patterns in Q which does not intersect $L(p)$, we have that when $l > \lceil S \rceil$, $Q_l = \emptyset$.

For Part (2), the same argument applies, with the difference that every pattern in $\mathbf{Q}_{\leq 1}$ is of length at most $\lceil S \rceil + 1$ instead of $\lceil S \rceil$. ∎

Corollary 27 *For every finite non-empty* $S \subseteq \Sigma^*$ *and* $k \in N^+$,
(1) $\mathcal{FC}(\mathcal{L}(\boldsymbol{P}_{\leq 1}), S, k, \lceil S \rceil + 2\min\{\lceil S \rceil, \lfloor \frac{k-1}{card(\Sigma)-1} \rfloor\}+2) \subseteq \mathcal{MINL}(\mathcal{L}(\boldsymbol{P}_{\leq 1}), S, k)$.
(2) $\mathcal{FC}(\mathcal{L}_\epsilon(\boldsymbol{P}_{\leq 1}), S, k, \lceil S \rceil + 2\min\{\lceil S \rceil+1, \lfloor \frac{k-1}{card(\Sigma)-1} \rfloor\}+2) \subseteq \mathcal{MINL}(\mathcal{L}_\epsilon(\boldsymbol{P}_{\leq 1}), S, k)$.

6 Concluding remarks

In scientific inference, parsimony of explanations is considered highly desirable, and its study under many different formulations (language and program size minimality *etc.*) has been carried out by many researchers [12,8]. The minimality we considered in this work is in terms of the number of elements generated by the hypotheses of up to a specified size. This notion of minimality is somewhat related to the language size, and we have shown how it can be related to the notion of minimality of subset relation proposed in [2] (MINL), in the cases where we can compute a sufficient length bound to the elements of the characteristic sets of the candidate hypotheses.

We mentioned related work [17,20] on such characteristic sets for some subclasses of the pattern languages, and also gave a few results. In summary, the results mentioned in this paper are for the following classes.

(I) the unions of the (erasing and non-erasing) regular pattern languages where the number of unions allowed is bounded by a factor of the alphabet size.

(II) the unbounded unions of the (erasing and non-erasing) substring and subsequence patterns.

(III) the bounded unions of the (erasing and non-erasing) languages generated by patterns which contain at most one variable occurrence.

While of little practical importance, the language class (III) is interesting to our study since it cannot be related to another notion of minimality based on syntactic construct (MMG) [4]. Below we list some classes of languages which we hope to see similar sufficient length bounds for the elements in their characteristic sets, if such bounds may be obtained.

(1) bounded unions of the (erasing and non-erasing) regular pattern languages where the bound is independent of the alphabet size. (That is, same as class (I) except that the alphabet size restriction on the number of unions is removed.)
(2) the unbounded unions of the (erasing and non-erasing) regular pattern languages of bounded segment length. [19]

References

[1] D. Angluin. Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21:46–62, 1980.

[2] D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135, 1980.

[3] D. Angluin. Inference of reversible languages. *Journal of the ACM*, 29:741–765, 1982.

[4] H. Arimura, T. Shinohara, and S. Otsuki. Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data. In *Proc. of the 11th Ann. Symp. on Theoretical Aspects of Comp. Sci. (STACS'94)*, volume 775, pages 649–660. Springer-Verlag, 1994.

[5] C. Chan, M. Garofalakis, and R. Rastogi. RE-tree: an efficient index structure for regular expressions. *The VLDB Journal*, 12(2):102–119, 2003.

[6] E. M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[7] M. Hirao, H. Hoshino, A. Shinohara, M. Takeda, and S. Arikawa. A practical algorithm to find the best subsequence patterns. In *Proceedings of the Third International Conference on Discovery Science*, volume 1967, pages 141–154, 2000.

[8] S. Jain, D. Osherson, J. Royer, and A. Sharma. *Systems that Learn: An Introduction to Learning Theory.* MIT Press, Cambridge, Mass., second edition, 1999.

[9] T. Jiang, E. Kinber, A. Salomaa, K. Salomaa, and S. Yu. Pattern languages with and without erasing. *International Journal Of Computer Mathematics*, 50:147–163, 1994.

[10] T. Jiang, A. Salomaa, K. Salomaa, and S. Yu. Decision problems for patterns. *Journal of Computer and System Sciences*, 50:53–63, 1995.

[11] S. Kobayashi and T. Yokomori. Identifiability of subspaces and homomorphic images of zero-reversible languages. In *Algorithmic Learning Theory: Eighth International Workshop (ALT '97)*, volume 1316, pages 48–61, 1997.

[12] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications.* Texts and Monographs in Computer Science. Springer-Verlag, 1993.

[13] T. Motoki, T. Shinohara, and K. Wright. The correct definition of finite elasticity: Corrigendum to identification of unions. In L. Valiant and M. Warmuth, editors, *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, page 375. Morgan Kaufmann, 1991.

[14] Y. K. Ng, H. Ono, and T. Shinohara. Measuring over-generalization in the minimal multiple generalizations of biosequences. In *Proceedings of The Eighth Conference on Discovery Science (DS'05)*, volume 3735, pages 176–188. Springer-Verlag, 2005.

[15] H. Ono and Y. K. Ng. Best fitting fixed-length substring patterns for a set of strings. In *Proceedings of The Eleventh International Computing and Combinatorics Conference (COCOON'05)*, volume 3595, pages 240–250. Springer-Verlag, 2005.

[16] M. Sato. Inductive inference of formal languages. *Bulletin of Informatics and Cybernetics*, 27(1):85–106, 1995.

[17] M. Sato, Y. Mukouchi, and D. Zheng. Characteristic sets for unions of regular pattern languages and compactness. In *Algorithmic Learning Theory: Ninth International Conference (ALT' 98)*, volume 1501, pages 220–233. Springer-Verlag, 1998.

[18] T. Shinohara. Polynomial time inference of extended regular pattern languages. In *RIMS Symposia on Software Science and Engineering, Kyoto, Japan*, volume 147, pages 115–127. Springer-Verlag, 1982.

[19] T. Shinohara and H. Arimura. Inductive inference of unbounded unions of pattern languages from positive data. *Theoretical Computer Science A*, 241:191–209, 2000.

[20] J. Uemura and M. Sato. Compactness and learning of classes of unions of erasing regular pattern languages. In *Algorithmic Learning Theory: Thirteenth International Conference (ALT' 02)*, volume 2533, pages 293–307. Springer-Verlag, 2002.

[21] K. Wright. Identification of unions of languages drawn from an identifiable class. In R. Rivest, D. Haussler, and M. Warmuth, editors, *Proceedings of the Second Annual Workshop on Computational Learning Theory*, pages 328–333. Morgan Kaufmann, 1989.