The Failure of the Strong Pumping Lemma for Multiple Context-Free Languages

Makoto Kanazawa, Gregory Kobele, Jens Michaelis, Sylvain Salvati, Ryo Yoshinaka

The final publication is available at www.springerlink.com.
DOI 10.1007/s00224-014-9534-z

The Failure of the Strong Pumping Lemma for Multiple Context-Free Languages

Makoto Kanazawa · **Gregory M. Kobele** · **Jens Michaelis** · **Sylvain Salvati** · **Ryo Yoshinaka**

Abstract Seki et al. (Theoretical Computer Science 88(2):191–229, 1991) showed that every m-multiple context-free language L is weakly $2m$-iterative in the sense that either L is finite or L contains a subset of the form $\{\, u_0 w_1^i u_1 \dots w_{2m}^i u_{2m} \mid i \in \mathbb{N} \,\}$, where $w_1 \dots w_{2n} \neq \varepsilon$. Whether every m-multiple context-free language L is $2m$-iterative, that is to say, whether all but finitely many elements z of L can be written as $z = u_0 w_1 u_1 \dots w_{2m} u_{2m}$ with $w_1 \dots w_{2m} \neq \varepsilon$ and $\{\, u_0 w_1^i u_1 \dots w_{2m}^i u_{2m} \mid i \in \mathbb{N} \,\} \subseteq L$, has been open. We show that there is a 3-multiple context-free language that is not k-iterative for any k.

Keywords Multiple context-free grammar · Pumping lemma

M. Kanazawa
National Institute of Informatics, 2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
E-mail: kanazawa@nii.ac.jp

G.M. Kobele
Computation Institute and Department of Linguistics, University of Chicago, Chicago, IL 60637, USA
E-mail: kobele@uchicago.edu

J. Michaelis
Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Postfach 10 01 31, D-33501 Bielefeld, Germany
E-mail: jens.michaelis@uni-bielefeld.de

S. Salvati
INRIA Bordeaux Sud-Ouest, LaBRI, 351, cours de la Libération, F-33405 Talence cedex, France
E-mail: sylvain.salvati@labri.fr

R. Yoshinaka
Graduate School of Informatics, Kyoto University, 36–1 Yoshida-Honmachi, Sakyo-ku, Kyoto, 606–8501, Japan
E-mail: ry@i.kyoto-u.ac.jp

1 Introduction

The study of iterative properties of the languages of *multiple context-free grammars* (MCFG) [14] has had a peculiar history.[1] Seki et al. [14] proved that any language L generated by an MCFG of dimension m (i.e., m-MCFG) is *weakly 2m-iterative* (in the sense of Greibach [3,2]): either L is finite or else it contains a subset of the form

$$\{\, u_0 w_1^i u_1 \dots w_{2m}^i u_{2m} \mid i \in \mathbb{N} \,\} \tag{1}$$

for some strings $u_0, u_1, \dots, u_{2m}$ and $w_1, \dots, w_{2m}$ such that $w_1 \dots w_{2m} \neq \varepsilon$.[2] Seki et al. [14] called this theorem a "pumping lemma" for m-MCFLs. Their proof of the theorem starts with an application of the pigeon-hole principle to a path in a derivation tree in a way familiar from the pumping lemma for context-free languages; beyond that, however, it involves much more intricate reasoning than in the context-free case, due to the complex relation between derivation trees of an MCFG and the derived strings. The proof goes roughly as follows.

Given a sufficiently long string z in the language L of an m-MCFG G, the derivation tree T for z must contain a "context" $U[]$ inside it that can be iterated any number of times.[3] That is to say, T can be written as $T = U'[U[T']]$, where $U[T']$ is a subtree of T which contains T' as a proper subtree, and for each $i \geq 0$, $U'[U^i[T']]$ is also a derivation tree. Here, the notation $U^i[T']$ is defined by

$$U^0[T'] = T',$$
$$U^{i+1}[T'] = U[U^i[T']].$$

In the case of a context-free grammar, each subtree of a derivation tree yields a single string. In the case of an m-MCFG, in contrast, each subtree of a derivation tree is associated with a tuple of strings. Thus, the contribution of the iterable context $U[]$ to the derived string is some function g mapping an n-tuple of strings to another n-tuple, for some $n \leq m$. Such a function can be specified by an equation of the form $g(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = (\alpha_1, \dots, \alpha_n)$ using variables $\boldsymbol{x}_i$ and strings α_i over $\Sigma \cup \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, where Σ is the terminal alphabet, such that each $\boldsymbol{x}_i$ occurs in a unique α_j. In the special case where $\alpha_j = w_{2j-1} \boldsymbol{x}_j w_{2j}$ for all $j = 1, \dots, n$ ($w_1, \dots, w_{2n} \in \Sigma^*$), iteration of $U[]$ inside the derivation tree translates into iteration of the strings $w_1, \dots, w_{2n}$ inside the derived string, giving rise to a set of the form (1). In general, since $\boldsymbol{x}_i$ may end up in some α_j with $j \neq i$, the effect of iterating $U[]$ in $T = U'[U[T']]$ is rather hard to describe. As a consequence, derivation trees of the form $U'[U^i[T']]$ do not (necessarily) generate a set of the form (1). One can see, however, that for large enough k, the k-fold composition g^k of g with itself has the property that if $g^k(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = (\beta_1, \dots, \beta_n)$, then for every $j = 1, \dots, n$,

---

[1] Around the same time as Kasami et al. [9] first introduced multiple context-free grammars, essentially the same formalism was proposed by Vijay-Shanker et al. [15] under the name *linear context-free rewriting systems* (LCFRS). In this paper, we mostly follow the terminology of Seki et al. [14].

[2] We let $\mathbb{N}$ denote the set of natural numbers $\{0, 1, 2, \dots\}$ and ε denote the empty string.

[3] Formally, a *context* is a tree with a single special leaf node ("hole"), which is labeled by $\square$. When $U[]$ is a context and T is a tree, $U[T]$ denotes the tree that results from removing the hole of $U[]$ and inserting T in its place.

β_j either is a constant string (i.e., string over Σ) or else contains $\boldsymbol{x}_j$. It follows that $g^{2k}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = g^k(\beta_1,\ldots,\beta_n) = (w_1\beta_1 w_2,\ldots,w_{2n-1}\beta_n w_{2n})$ for some constant strings $w_1,\ldots,w_{2n}$ such that $w_{2j-1}w_{2j} = \varepsilon$ whenever β_j is a constant string. It is not difficult to see that this implies that $g^{(i+1)k}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = (w_1^i\beta_1 w_2^i,\ldots,w_{2n-1}^i\beta_n w_{2n}^i)$. Thus, derivation trees $U'[U^{(i+1)k}[T']]$ $(i \geq 0)$ yield a subset of L of the required form (1). Crucially, the original string z is not an element of this set.

By a strange quirk of fate, this proof was erroneously claimed by Radzinski [13] to implicitly demonstrate a much stronger property,[4] namely, that every m-MCFL L is $2m$*-iterative* (in the sense of Greibach [3]): all but finitely many $z \in L$ can be written as $z = u_0 w_1 u_1 \ldots w_{2m} u_{2m}$ such that $w_1 \ldots w_{2m} \neq \varepsilon$ and $\{\, u_0 w_1^i u_1 \ldots w_{2m}^i u_{2m} \mid i \in \mathbb{N} \,\} \subseteq L$. More strangely, Groenink [5] just took Radzinski's word for it (see also [4]). A more recent book by Kracht [10] also states this property as a theorem.

We refer to the assertion that every m-MCFL is $2m$-iterative as the *strong pumping lemma for m-MCFLs*, to distinguish it from Seki et al.'s [14] theorem. It is clear that no simple modification of the method of Seki et al. can establish the strong pumping lemma for m-MCFLs. It is only when the iterable context $U[]$ maps an n-tuple $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ to an n-tuple of the form $(w_1\boldsymbol{x}_1 w_2,\ldots,w_{2n-1}\boldsymbol{x}_n w_{2n})$ that it is possible to conclude, analogously to the context-free case, that the given string z contains factors $w_1,\ldots,w_{2m}$ that can be pumped up and down without pushing the resulting string outside of the given m-MCFL.[5] Kanazawa [6] called such a well-behaved iterable context an *even pump* in his proof that an m-MCFG satisfying the condition of *well-nestedness* always generates a $2m$-iterative set. This proof works by induction on m. The base case is handled by the fact that well-nested 1-MCFGs are just CFGs. For the induction step, Kanazawa showed that given a well-nested m-MCFG G, one can always find a well-nested $(m-1)$-MCFG G' for the language L' consisting of strings generated by G with derivation trees containing no even pump. Hence the language L of G is a union of some $2m$-iterative set and L', which, by induction hypothesis, is a $2(m-1)$-iterative set. It follows that L is $2m$-iterative, completing the induction. This method is such that derivation trees of G' have very different shapes from the original derivation trees of G for the same strings. Whereas the method also works for 2-MCFGs in general, the well-nestedness property is essential for $m \geq 3$, and there is no obvious way of extending it to the non-well-nested case.

In this paper, we prove that the strong pumping lemma indeed fails for non-well-nested m-MCFGs for $m \geq 3$. We do so by exhibiting a particular 3-MCFG that generates a language that is non-iterative in a very strong sense. This language, which we call H, is not k-iterative for any k. It is not even *finitely pumpable* in the sense of Groenink [5,4], a condition which is similar to k-iterativity but allows the number of iterable factors to vary from string to string. In fact, H contains an infinite subset $\{\, v_n \mid n \in \mathbb{N} \,\}$ consisting of strings that are *almost anti-iterative* in the following sense: whenever $v_n = u_0 w_1 u_1 \ldots w_k u_k$ and $w_1 \ldots w_k \neq \varepsilon$ (for any k), it holds that

$$|\{\, i \mid i > 1 \text{ and } u_0 w_1^i u_1 \ldots w_k^i u_k \in H \,\}| \leq 1.$$

[4] See footnote 10 of Radzinski [13]. Radzinski refers to the technical report [9] rather than the journal article [14] based on it, but the proof is the same in both papers.

[5] A string v is a *factor* of a string z if $z = uvw$ for some strings u, w.

Most of the rest of the paper is devoted to the proof of this property of the language H (section 3). Before we get to it, we briefly review basic notions concerning multiple context-free grammars for readers unfamiliar with this grammar formalism (section 2). The proof in section 3 does not use any general properties of MCFLs, and can be followed by anyone who understands the definition of the language H.

2 Multiple Context-Free Grammars

Like a context-free grammar, a *multiple context-free grammar* is a quadruple $G = (N, \Sigma, P, S)$, where N is a finite set of nonterminals, Σ is a finite set of terminals, P is a set of rules, and S is a designated nonterminal. While a nonterminal of a CFG is associated with a set of terminal strings, a nonterminal of an MCFG is interpreted as a q-ary relation on terminal strings, where q is the *dimension* of the nonterminal. Each nonterminal comes with a unique dimension. (So the set N can be thought of as a *ranked alphabet*.) The dimension of the designated nonterminal S is always 1. A rule is of the form

$$A(\alpha_1, \dots, \alpha_q) \leftarrow B_1(\boldsymbol{x}_{1,1}, \dots, \boldsymbol{x}_{1,q_1}), \dots, B_n(\boldsymbol{x}_{n,1}, \dots, \boldsymbol{x}_{n,q_n}),$$

where $n \geq 0$, $A, B_1, \dots, B_n$ are nonterminals of dimension $q, q_1, \dots, q_n$, respectively, the $\boldsymbol{x}_{i,j}$ are pairwise distinct *variables*, which are symbols not in Σ, and $\alpha_1, \dots, \alpha_q$ are strings over $\Sigma \cup \{ \boldsymbol{x}_{i,j} \mid 1 \leq i \leq n, 1 \leq j \leq q_i \}$ such that each $\boldsymbol{x}_{i,j}$ occurs at most once in $\alpha_1 \dots \alpha_q$.

A rule is interpreted like a universally quantified implication from right to left. Define a predicate $\vdash_G$ that holds of expressions of the form $A(u_1, \dots, u_q)$ (called *facts*) inductively as follows:

- If $A(u_1, \dots, u_q) \leftarrow$ is a rule of G, then $\vdash_G A(u_1, \dots, u_q)$.
- If $A(\alpha_1, \dots, \alpha_q) \leftarrow B_1(\boldsymbol{x}_{1,1}, \dots, \boldsymbol{x}_{1,q_1}), \dots, B_n(\boldsymbol{x}_{n,1}, \dots, \boldsymbol{x}_{n,q_n})$ is a rule of G and $\vdash_G B_i(w_{i,1}, \dots, w_{i,q_i})$ for $i = 1, \dots, n$, then $\vdash_G A(u_1, \dots, u_q)$, where $(u_1, \dots, u_q)$ is the result of substituting $w_{i,j}$ for each $\boldsymbol{x}_{i,j}$ in $(\alpha_1, \dots, \alpha_q)$.

When $\vdash_G A(u_1, \dots, u_q)$, we say that $A(u_1, \dots, u_q)$ is *derivable* (in G). (We sometimes write $\vdash$ instead of $\vdash_G$ when the grammar is clear from the context.) The language of G is defined by $L(G) = \{ w \in \Sigma^* \mid \vdash_G S(w) \}$.

An MCFG is an *m-MCFG* if the dimension of nonterminals does not exceed m. The language of an m-MCFG is called an *m-MCFL*. It is shown by Seki et al. [14] that each m-MCFG has an equivalent one such that the variables on the right-hand side of any rule all appear in the left-hand side. Such an MCFG is called *non-deleting*.

A rule $A(\alpha_1, \dots, \alpha_q) \leftarrow B_1(\boldsymbol{x}_{1,1}, \dots, \boldsymbol{x}_{1,q_1}), \dots, B_n(\boldsymbol{x}_{n,1}, \dots, \boldsymbol{x}_{n,q_n})$ is called *non-permuting* if for each $i = 1, \dots, n$ and each j, k such that $1 \leq j < k \leq q_i$, it is not the case that

$$\varphi(\alpha_1 \dots \alpha_q) = \boldsymbol{x}_{i,k} \boldsymbol{x}_{i,j},$$

where φ is the homomorphism that erases all symbols in Σ and all variables other than $\boldsymbol{x}_{i,j}$ and $\boldsymbol{x}_{i,k}$. An MCFG G is called non-permuting if all its rules are non-permuting. Every m-MCFG has an equivalent non-deleting non-permuting m-MCFG [11, 10].

A non-deleting non-permuting MCFG is called *well-nested* if every rule $A(\alpha_1,\dots,\alpha_q) \leftarrow B_1(\boldsymbol{x}_{1,1},\dots,\boldsymbol{x}_{1,q_1}),\dots,B_n(\boldsymbol{x}_{n,q},\dots,\boldsymbol{x}_{n,q_n})$ satisfies the following condition: whenever $i \neq i'$, $1 \leq j < k \leq q_i$, $1 \leq j' < k' \leq q_{i'}$, it is not the case that

$$\chi(\alpha_1 \dots \alpha_q) = \boldsymbol{x}_{i,j}\boldsymbol{x}_{i',j'}\boldsymbol{x}_{i,k}\boldsymbol{x}_{i',k'}$$

where χ is the homomorphism that erases all symbols in Σ and all variables other than $\boldsymbol{x}_{i,j},\boldsymbol{x}_{i,k},\boldsymbol{x}_{i',j'},\boldsymbol{x}_{i',k'}$. Kanazawa [6] showed that the languages of well-nested m-MCFGs are all $2m$-iterative. See also [8] for the effect of the well-nestedness condition on the generative power of MCFGs.

In order to rigorously define the notion of a derivation tree, we view the rule set P as a ranked alphabet where $\pi \in P$ has rank n if the right-hand side of π has n occurrences of nonterminals. A derivation tree of $G = (N,\Sigma,P,S)$ is a local set of trees over P, defined inductively as follows:

- If $\pi = A(u_1,\dots,u_q) \leftarrow$ is a rule in P, then π is a derivation tree for $A(u_1,\dots,u_q)$.
- If $\pi = A(\alpha_1,\dots,\alpha_q) \leftarrow B_1(\boldsymbol{x}_{1,1},\dots,\boldsymbol{x}_{1,q_1}),\dots,B_n(\boldsymbol{x}_{n,1},\dots,\boldsymbol{x}_{n,q_n})$ is a rule in P and for $i = 1,\dots,n$, T_i is a derivation tree for $B_i(w_{i,1},\dots,w_{i,q_i})$, then $\pi T_1 \dots T_n$ is a derivation tree for $A(u_1,\dots,u_q)$, where $(u_1,\dots,u_q)$ is the result of substituting $w_{i,j}$ for each $\boldsymbol{x}_{i,j}$ in $(\alpha_1,\dots,\alpha_q)$.

A derivation tree for $A(u_1,\dots,u_q)$ is a derivation tree *of type A*. A *complete* derivation tree is a derivation tree of type S, and it is said to be a derivation tree for w if it is a derivation tree for $S(w)$. When T is a derivation tree for a fact $A(u_1,\dots,u_q)$, we also say T *derives* $A(u_1,\dots,u_q)$. Clearly, $\vdash_G A(u_1,\dots,u_q)$ holds if and only if G has a derivation tree that derives $A(u_1,\dots,u_q)$.

When a derivation tree of type B contains a derivation tree of type A as a subtree, the result of replacing that subtree by any other derivation tree of type A is again a derivation tree of type B. When a complete derivation tree T for w has a path containing more nodes than the number of nonterminals, then there must be a nonterminal A and two nodes on that path such that the subtree rooted at each of the two nodes is a derivation tree of type A. This is the starting point of Seki et al.'s [14] proof of their pumping lemma.

Example 1 Consider the following 2-MCFG:

$$\begin{aligned}
&\pi_1 : S(\boldsymbol{x}_1\#\boldsymbol{x}_2) \leftarrow D(\boldsymbol{x}_1,\boldsymbol{x}_2)\\
&\pi_2 : D(\varepsilon,\varepsilon) \leftarrow\\
&\pi_3 : D(\boldsymbol{x}_1\boldsymbol{y}_1,\boldsymbol{x}_2\boldsymbol{y}_2) \leftarrow E(\boldsymbol{x}_1,\boldsymbol{x}_2),D(\boldsymbol{y}_1,\boldsymbol{y}_2)\\
&\pi_4 : E(c\boldsymbol{x}_1\bar{c},c\boldsymbol{x}_2\bar{c}) \leftarrow D(\boldsymbol{x}_1,\boldsymbol{x}_2)
\end{aligned}$$

(π_1,π_2,π_3,π_4 are the names of the rules.) Here, S is the designated nonterminal, and all other nonterminals are of rank 2. This grammar generates $\{\, w\#w \mid w \in D_1^* \,\}$, where D_1^* is the Dyck language over the alphabet $\{c,\bar{c}\}$. Note that the third rule is not well-nested. Figure 1 shows a derivation tree for $cc\bar{c}\bar{c}c\bar{c}\#cc\bar{c}\bar{c}c\bar{c}$, alongside of the same tree with each node annotated by the fact derived by the subtree rooted at that node.

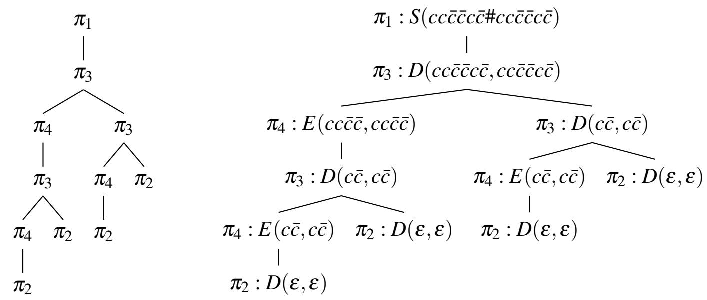

Fig. 1 A derivation tree for $cc\bar{c}\bar{c}c\bar{c}\#cc\bar{c}\bar{c}c\bar{c}$ (left) and the same tree augmented with additional information about what fact is derived at each step (right).

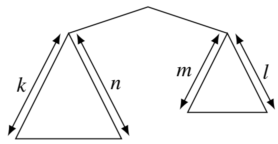

Fig. 2 Derivation tree for $J(a^{k+1}, a^m cv\bar{c}dw\bar{d}b^n, b^{l+1})$.

3 Counterexample to the Strong Pumping Lemma for 3-MCFLs

We fix two alphabets:

$$\Sigma = \{c, \bar{c}, d, \bar{d}\},$$
$$\hat{\Sigma} = \Sigma \cup \{a, b\}.$$

Define a 3-MCFL $H \subseteq \hat{\Sigma}^*$ by the following 3-MCFG, where we use the symbol H itself as the designated nonterminal:

$$\begin{aligned} &H(\boldsymbol{x}_2) \leftarrow J(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \\ &J(a\boldsymbol{x}_1, \boldsymbol{y}_1 c\boldsymbol{x}_2 \bar{c} d\boldsymbol{y}_2 \bar{d}\boldsymbol{x}_3, \boldsymbol{y}_3 b) \leftarrow J(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3), J(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3) \\ &J(a, \varepsilon, b) \leftarrow \end{aligned}$$

This is our counterexample to the strong pumping lemma. Note that the second rule is not well-nested. When $J(u_1, u_2, u_3)$ is derivable in this grammar, we always have $u_1 = a^{k+1}$, $u_3 = b^{l+1}$ for some $k, l \in \mathbb{N}$, and u_2 is either ε or a string of the form $a^m cv\bar{c}dw\bar{d}b^n$ for some $v, w \in H$ and $m, n \geq 1$. In the latter case, the (unique) derivation tree for $J(a^{k+1}, a^m cv\bar{c}dw\bar{d}b^n, b^{l+1})$ is a binary tree T where k and n are the numbers of nodes on the leftmost and rightmost branches, respectively, of the left immediate subtree of T, and m and l are the numbers of nodes on the leftmost and rightmost branches, respectively, of the right immediate subtree of T (Figure 2).

The language H is related to a context-free language over Σ via the homomorphism $\psi \colon \hat{\Sigma}^* \to \Sigma^*$ defined by:

$$\psi(e) = \begin{cases} \varepsilon & \text{if } e \in \{a, b\}, \\ e & \text{if } e \in \Sigma. \end{cases}$$

It is easy to see that $\psi(H)$ is a context-free language included in the Dyck language D_2^* over the alphabet Σ, where $(c,\bar{c})$ and $(d,\bar{d})$ are each regarded as a matching pair of parentheses. The homomorphism ψ is an injection when restricted to the strings in H, and for each $v \in H$, $\psi(v)$ encodes in an obvious way the unique derivation tree for v. We can learn a lot about iterative properties of the 3-MCFL H from the CFL $\psi(H)$, so we begin by studying the latter.

3.1 Properties of the CFL $V = \psi(H)$

The goal of this section is to state a necessary condition for $w \in \Sigma^+$ to be in

$$\{\, w \mid ww \text{ is a factor of some string in } \psi(H) \,\}.$$

In what follows, we use regular expressions and (recursive) equations involving regular expressions to define various languages. In regular expressions, the vertical bar "|" denotes union, and is assumed to have lower precedence than all other operators.

Define the *reduction* relation $\rhd \in \Sigma^* \times \Sigma^*$ by

$$\rhd = \{\, (v_1 c\bar{c} v_2, v_1 v_2) \mid v_1, v_2 \in \Sigma^* \,\} \cup \{\, (v_1 d\bar{d} v_2, v_1 v_2) \mid v_1, v_2 \in \Sigma^* \,\}.$$

We write $\rhd^*$ for the reflexive transitive closure of the relation $\rhd$, and $\rhd^n$ for the n-fold composition of $\rhd$ with itself (more precisely, $\rhd^{n+1}$ is $\rhd$ composed with $\rhd^n$, where $\rhd^0$ is the identity relation). When $v \rhd^* w$, we say *v reduces to w*, and when $v \rhd^n w$, we say *v reduces to w in n steps*. A string $w \in \Sigma^*$ is said to be in *normal form* if neither $c\bar{c}$ nor $d\bar{d}$ is a factor of w. It is well known that the relation $\rhd^*$ has the confluence (i.e., Church-Rosser) property and each string $w \in \Sigma^*$ reduces to a unique string in normal form, which is called the *normal form* of w. We write $\mathrm{nf}(w)$ for the normal form of w. The *Dyck language* D_2^* over Σ is defined as $D_2^* = \{\, w \in \Sigma^* \mid \mathrm{nf}(w) = \varepsilon \,\}$.

Lemma 2 *The following conditions hold of all $u, v, w, v' \in \Sigma^*$:*

(i) *If $v \rhd^* v' \in \bar{c}\Sigma^*$, then $\mathrm{nf}(vw) \in \bar{c}\Sigma^*$.*
(ii) *If $v \rhd^* v' \in \bar{d}\Sigma^*$, then $\mathrm{nf}(vw) \in \bar{d}\Sigma^*$.*
(iii) *If $v \rhd^* v' \in \Sigma^* c$, then $\mathrm{nf}(uv) \in \Sigma^* c$.*
(iv) *If $v \rhd^* v' \in \Sigma^* d$, then $\mathrm{nf}(uv) \in \Sigma^* d$.*
(v) *If $v \rhd^* v' \in \Sigma^* c\bar{d}\Sigma^*$, then $\mathrm{nf}(uvw) \in \Sigma^* c\bar{d}\Sigma^*$.*
(vi) *If $v \rhd^* v' \in \Sigma^* d\bar{c}\Sigma^*$, then $\mathrm{nf}(uvw) \in \Sigma^* d\bar{c}\Sigma^*$.*

Proof (i). Since $v \rhd^* v' \in \bar{c}\Sigma^*$ implies $vw \rhd^* v'w \in \bar{c}\Sigma^*$ and, by the confluence property, $\mathrm{nf}(vw) = \mathrm{nf}(v'w)$, it suffices to show that $z \in \bar{c}\Sigma^*$ implies $\mathrm{nf}(z) \in \bar{c}\Sigma^*$ for all $z \in \Sigma^*$. We prove this by induction on the number of reduction steps from z to $\mathrm{nf}(z)$. Suppose $z = \bar{c}y$. If $z = \mathrm{nf}(z)$, then $\mathrm{nf}(z) \in \bar{c}\Sigma^*$. Otherwise, $z = \bar{c}y \rhd^n \mathrm{nf}(z)$ for some $n \geq 1$. Then $\bar{c}y \rhd x \rhd^{n-1} \mathrm{nf}(z) = \mathrm{nf}(x)$ for some $x \in \bar{c}\Sigma^*$. By the induction hypothesis applied to x, we obtain $\mathrm{nf}(z) \in \bar{c}\Sigma^*$.

Part (ii)–(vi) may be proved similarly. □

Lemma 3 *Let $w \in \Sigma^*$ and suppose $\mathrm{nf}(w) = e_1 \dots e_n$ for some $e_1, \dots, e_n \in \Sigma$. Then there exist $u_0, \dots, u_n \in \Sigma^*$ such that $w = u_0 e_1 u_1 \dots e_n u_n$ and $\mathrm{nf}(u_i) = \varepsilon$ for $i = 0, \dots, n$.*

Proof By induction on the number of reduction steps from w to $e_1 \dots e_n$. □

If K is a set of strings, let $\mathrm{fac}(K)$ be the set of factors of elements of K, i.e.,

$$\mathrm{fac}(K) = \{ v \mid uvw \in K \}.$$

Since the relation "is a factor of" is reflexive and transitive, $\mathrm{fac}(\mathrm{fac}(K)) = \mathrm{fac}(K)$ always holds.

Lemma 4 *For every $w \in \mathrm{fac}(D_2^*)$, it holds that* $\mathrm{nf}(w) \in (\bar{c} \mid \bar{d})^*(c \mid d)^*$.

Proof By the definition of normal form, $\mathrm{nf}(w)$ cannot contain $c\bar{c}$ or $d\bar{d}$ as a factor. Now $\mathrm{nf}(w)$ cannot contain $c\bar{d}$ or $d\bar{c}$ as a factor, either. To see this, let $uwv \in D_2^*$ and suppose $c\bar{d}$ or $d\bar{c}$ is a factor of $\mathrm{nf}(w)$. Then by Lemma 2, part (v) and (vi), $\mathrm{nf}(uwv)$ contains $c\bar{d}$ or $d\bar{c}$ as a factor, contradicting $\mathrm{nf}(uwv) = \varepsilon$. The desired conclusion now follows easily. □

Lemma 5 *If $vw \in D_2^*$, then* $\mathrm{nf}(v) \in (c \mid d)^*$ *and* $\mathrm{nf}(w) \in (\bar{c} \mid \bar{d})^*$.

Proof Suppose $vw \in D_2^*$. By Lemma 4, $\mathrm{nf}(v)$ and $\mathrm{nf}(w)$ both belong to $(\bar{c} \mid \bar{d})^*(c \mid d)^*$. If $\mathrm{nf}(v) \in (\bar{c} \mid \bar{d})^+(c \mid d)^*$, then by Lemma 2, part (i) and (ii), $\mathrm{nf}(vw) \in (\bar{c} \mid \bar{d})\Sigma^*$, contradicting $vw \in D_2^*$. Hence $\mathrm{nf}(v) \in (c \mid d)^*$. Similarly, we can conclude $\mathrm{nf}(w) \in (\bar{c} \mid \bar{d})^*$ using Lemma 2, part (iii) and (iv). □

The set D_2 of *Dyck primes* over Σ is defined as $D_2 = cD_2^*\bar{c} \mid dD_2^*\bar{d}$. It is well known and easy to see that D_2^* indeed equals $(D_2)^*$.

Define context-free languages V, L, R by[6]

$$\begin{aligned} V &= \varepsilon \mid LR, \\ L &= cV\bar{c}, \\ R &= dV\bar{d}. \end{aligned}$$

Then it is easy to see that $V \subset D_2^*$, $L \subset D_2$, $R \subset D_2$.

Lemma 6 $\mathrm{fac}(V) \cap \Sigma^2 = \{cc, c\bar{c}, \bar{c}d, dc, d\bar{d}, \bar{d}\bar{c}, \bar{d}\bar{d}\}$.

Proof First, note that $V = \varepsilon \mid LR$ implies that every $v \in V$ satisfies $v \in \varepsilon \mid c\Sigma^*\bar{d}$. Let F be the set on the right-hand side of the equation to be proved. We can show by induction on the length of v that $v \in V$ and $w \in \mathrm{fac}(v) \cap \Sigma^2$ imply $w \in F$. Suppose $v \in V$ and $w \in \mathrm{fac}(v) \cap \Sigma^2$. Then $v \in LR = cV\bar{c}dV\bar{d}$, so $v = cv_1\bar{c}dv_2\bar{d}$ for some $v_1, v_2 \in V$. Hence either $w \in \mathrm{fac}(\{v_1, v_2\}) \cap \Sigma^2$ or $w \in \{cc, c\bar{c}, \bar{d}\bar{c}, \bar{c}d, dc, d\bar{d}, \bar{d}\bar{d}\} = F$. By induction hypothesis, $\mathrm{fac}(\{v_1, v_2\}) \cap \Sigma^2 \subseteq F$, so it follows that $w \in F$. This establishes $\mathrm{fac}(V) \cap \Sigma^2 \subseteq F$. To see the converse inclusion, just note that for $u = cc\bar{c}dd\bar{d}\bar{c}dc\bar{c}dd\bar{d}\bar{d} \in V$, we have $\mathrm{fac}(u) \cap \Sigma^2 = F$. □

Lemma 7 $V = \psi(H)$.

[6] As usual, the sets V, L, R are understood to be the components of the least solution to these equations.

Proof Applying the homomorphism ψ in each rule of the 3-MCFG for H, we get

$$H(\boldsymbol{x}_2) \leftarrow J(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$$
$$J(\boldsymbol{x}_1, \boldsymbol{y}_1 c \boldsymbol{x}_2 \bar{c} d \boldsymbol{y}_2 \bar{d} \boldsymbol{x}_3, \boldsymbol{y}_3) \leftarrow J(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3), J(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3)$$
$$J(\varepsilon, \varepsilon, \varepsilon) \leftarrow$$

In this grammar, whenever $J(u_1, u_2, u_3)$ is derivable, $u_1 = u_3 = \varepsilon$. So the first and third arguments of J can be dropped, and the grammar can be simplified to

$$J(c\boldsymbol{x}\bar{c}d\boldsymbol{y}\bar{d}) \leftarrow J(\boldsymbol{x}), J(\boldsymbol{y})$$
$$J(\varepsilon) \leftarrow$$

This is just a context-free grammar for V. □

Lemma 8 $D_2 \cap \mathrm{fac}(V) = L \mid R$.

Proof Since $V = \varepsilon \mid LR$ and $L \mid R \subseteq D_2$, it is clear that $L \mid R \subseteq D_2 \cap \mathrm{fac}(V)$.

For the converse inclusion, we prove by induction on the length of $x \in V$ that $x = uvw$ and $v \in D_2$ implies $v \in L \mid R$. The base case of $x = \varepsilon$ is trivial. For the induction step, let $x = cy\bar{c}dz\bar{d}$, where $y, z \in V$, and suppose $x = uvw$ and $v \in D_2$. We distinguish three cases.

Case 1. v is a factor of $cy\bar{c}$. If $v = cy\bar{c}$, then $v \in L$, and if v is a factor of y, then $v \in L \mid R$ by the induction hypothesis. If $v = cy'$, where y' is a prefix of y, then $\mathrm{nf}(v) = \mathrm{nf}(cy') \in c(c \mid d)^*$ by Lemma 5. So $\mathrm{nf}(v) \neq \varepsilon$, contradicting $v \in D_2$. Likewise, if $v = y''\bar{c}$, where y'' is a suffix of y, then $\mathrm{nf}(v) = \mathrm{nf}(y''\bar{c}) \in (\bar{c} \mid \bar{d})^*\bar{c}$ and $\mathrm{nf}(v) \neq \varepsilon$, contradicting $v \in D_2$.
Case 2. v is a factor of $dz\bar{d}$. This case is completely analogous to Case 1, and we can conclude $v \in L \mid R$.
Case 3. $v = v'v''$, where v' is a non-empty suffix of $cy\bar{c}$ and v'' is a non-empty prefix of $dz\bar{d}$. Since $v \in D_2$, v cannot equal $x = cy\bar{c}dz\bar{d}$. So either v' is a suffix of $y\bar{c}$, in which case $\mathrm{nf}(v) = \mathrm{nf}(v'v'') \in (\bar{c} \mid \bar{d})^*\bar{c}(c \mid d)^*$ by Lemma 5, or else v'' is a prefix of dz, in which case $\mathrm{nf}(v) = \mathrm{nf}(v'v'') \in (\bar{c} \mid \bar{d})^*d(c \mid d)^*$, again by Lemma 5. In either case, $\mathrm{nf}(v) \neq \varepsilon$, contradicting $v \in D_2$.

We have seen that $v \in L \mid R$ holds in all cases, and the induction step is complete. □

Lemma 9 $D_2^* \cap \mathrm{fac}(V) = V \mid L \mid R$.

Proof Since $V = \varepsilon \mid LR$ and $L \mid R \subseteq D_2$, it is clear that $V \mid L \mid R \subseteq D_2^* \cap \mathrm{fac}(V)$.

For the converse inclusion, suppose $w \in D_2^* \cap \mathrm{fac}(V)$. Since any factor of a string in $\mathrm{fac}(V)$ is itself in $\mathrm{fac}(V)$, it follows that $w \in (D_2 \cap \mathrm{fac}(V))^*$. By Lemma 8, $w \in (L \mid R)^* \cap \mathrm{fac}(V)$. Since any string in $LL \mid RL \mid RR$ has one of $\bar{c}c, \bar{d}c, \bar{d}d$ as a factor, Lemma 6 implies $(LL \mid RL \mid RR) \cap \mathrm{fac}(V) = \varnothing$. It follows that $(L \mid R)^2 \cap \mathrm{fac}(V) = LR$ and for $n \geq 3$,

$$\begin{aligned}(L \mid R)^n \cap \mathrm{fac}(V) &= ((L \mid R)^2 \cap \mathrm{fac}(V))(L \mid R)^{n-2} \cap \mathrm{fac}(V) \\ &= LR(L \mid R)^{n-2} \cap \mathrm{fac}(V) \\ &\subseteq L((RL \mid RR) \cap \mathrm{fac}(V))(L \mid R)^{n-3} \\ &= \varnothing.\end{aligned}$$

So

$$\begin{aligned} w &\in (\varepsilon \mid (L \mid R) \mid (L \mid R)^2) \cap \mathrm{fac}(V) \\ &= \varepsilon \mid (L \mid R) \mid LR \\ &= V \mid L \mid R. \end{aligned}$$

This proves $D_2^* \cap \mathrm{fac}(V) \subseteq V \mid L \mid R$. □

Lemma 10 *Let $u, w \in \Sigma^*$ and $v \in \Sigma^+$. If $uv \in V$ and $vw \in V$, then $u = w = \varepsilon$.*

Proof Since $V \subset D_2^*$, Lemma 5 implies $\mathrm{nf}(v) \in (\bar{c} \mid \bar{d})^* \cap (c \mid d)^*$, and hence $\mathrm{nf}(v) = \varepsilon$. It follows that $\mathrm{nf}(u) = \mathrm{nf}(w) = \varepsilon$, too, and hence u, v, w are all in D_2^*. By Lemma 9, u, v, w are all in $V \mid L \mid R$. Since $v \neq \varepsilon$, the strings uv and vw are both in $V - \{\varepsilon\} = LR = cV\bar{c}dV\bar{d}$. So v ends in $\bar{d}$ and begins in c. If $u \neq \varepsilon$, then $u \in LR \mid L \mid R$, so $u \in \Sigma^*(\bar{c} \mid \bar{d})$. This implies either $\bar{c}c$ or $\bar{d}c$ is a factor of $uv \in V$, contradicting Lemma 6. Therefore, $u = \varepsilon$. Similarly, we can use Lemma 6 to conclude $w = \varepsilon$. □

We say that a string u is a *proper prefix* (*proper suffix*) of a string v if u is a prefix (suffix) of v and $u \neq v$. Lemma 10 implies that no proper prefix or proper suffix of a string in V can belong to V, which is to say that V is both *prefix-free* and *suffix-free*.

Lemma 11

$$\begin{aligned} \mathrm{fac}(V) \subseteq\ & (V \mid L \mid R) \mid \\ & (V \mid R)(\bar{c}R \mid \bar{d})^*(\bar{c} \mid \bar{c}R \mid \bar{d}) \mid \\ & (c \mid Ld \mid d)(c \mid Ld)^*(V \mid L) \mid \\ & (V \mid R)(\bar{c}R \mid \bar{d})^*\bar{c}d(c \mid Ld)^*(V \mid L). \end{aligned}$$

Proof Suppose $w \in \mathrm{fac}(V)$. By Lemma 4, $\mathrm{nf}(w) \in (\bar{c} \mid \bar{d})^m(c \mid d)^n$ for some $m, n \geq 0$, and by Lemma 3, there are strings $u_0, \dots, u_{m+n}$ such that $\mathrm{nf}(u_i) = \varepsilon$ for each $i = 0, \dots, m+n$ and

$$w \in u_0(\bar{c} \mid \bar{d})u_1 \dots (\bar{c} \mid \bar{d})u_m(c \mid d)u_{m+1} \dots (c \mid d)u_{m+n}.$$

Since u_i is a factor of $w \in \mathrm{fac}(V)$, $u_i \in D_2^* \cap \mathrm{fac}(V)$. Lemma 9 then implies $u_i \in V \mid L \mid R$.

By Lemma 6, each of the following sets is disjoint from $\mathrm{fac}(V)$:

$$\begin{aligned} &\bar{c}(\bar{c} \mid \bar{d}), \qquad (c \mid d)d, \\ &\bar{d}(c \mid d), \qquad (\bar{c} \mid \bar{d})c. \end{aligned}$$

This implies that the following conditions hold:

$$u_0 \in V \mid R \text{ if } m \geq 1, \tag{2}$$

$$u_{m+n} \in V \mid L \text{ if } n \geq 1, \tag{3}$$

$$u_i \in \varepsilon \mid R \text{ if } u_i \text{ is preceded by } \bar{c}, \tag{4}$$

$$u_i \in R \text{ if } u_i \text{ is preceded by } \bar{c} \text{ and is followed by } \bar{c} \text{ or } \bar{d}, \tag{5}$$

$$u_i = \varepsilon \text{ if } u_i \text{ is preceded by } \bar{d}, \tag{6}$$

$$u_i = \varepsilon \text{ if } u_i \text{ is followed by } c, \tag{7}$$

$$u_i \in \varepsilon \mid L \text{ if } u_i \text{ is followed by } d, \tag{8}$$

$$u_i \in L \text{ if } u_i \text{ is preceded by } c \text{ or } d \text{ and is followed by } d. \tag{9}$$

Case 1. $m = n = 0$. Then $w = u_0 \in V \mid L \mid R$.
Case 2. $m \geq 1, n = 0$. Then $w \in u_0(\bar{c} \mid \bar{d})u_1 \dots (\bar{c} \mid \bar{d})u_m$. By (2), (4), (5), and (6), we get $w \in (V \mid R)(\bar{c}R \mid \bar{d})^*(\bar{c} \mid \bar{c}R \mid \bar{d})$.
Case 3. $m = 0, n \geq 1$. Then $w \in u_0(c \mid d) \dots u_{n-1}(c \mid d)u_n$. By (3), (7), (8), and (9), we get $w \in (c \mid Ld \mid d)(c \mid Ld)^*(V \mid L)$.
Case 4. $m, n \geq 1$. By (4), (6), (7), and (8), we see that $u_m = \varepsilon$. Since $(\bar{c}c \mid \bar{d}c \mid \bar{d}d) \cap \mathrm{fac}(V) = \varnothing$,

$$w \in u_0(\bar{c} \mid \bar{d})u_1 \dots (\bar{d} \mid \bar{d})u_{m-1}\bar{c}du_{m+1}(c \mid d) \dots u_{m+n-1}(c \mid d)u_{m+n}.$$

By (2), (3), (5), (6), (7), and (9), we see that $w \in (V \mid R)(\bar{c}R \mid \bar{d})^*\bar{c}d(c \mid Ld)^*(V \mid L)$.

This proves the lemma. □

Lemma 12 *If $w \in \Sigma^+$ and $ww \in \mathrm{fac}(V)$, then one of the following conditions holds:*

(i) $w \in (\bar{c}R \mid \bar{d})^+$.
(ii) $w \in R(\bar{c}R \mid \bar{d})^*\bar{c}$.
(iii) $w \in (c \mid Ld)^+$.
(iv) $w \in d(c \mid Ld)^*L$.
(v) $w \in (V \mid R)(\bar{c}R \mid \bar{d})^m\bar{c}d(c \mid Ld)^n(V \mid L)$ *for some $m, n \geq 0$ such that $m \neq n$.*

Proof Suppose $w \neq \varepsilon$ and $ww \in \mathrm{fac}(V)$. Since $w \in \mathrm{fac}(V)$, by Lemma 11,

$$\begin{aligned}\mathrm{fac}(V) \subseteq\ & (V \mid L \mid R) \mid \\ & (V \mid R)(\bar{c}R \mid \bar{d})^*(\bar{c} \mid \bar{c}R \mid \bar{d}) \mid \\ & (c \mid Ld \mid d)(c \mid Ld)^*(V \mid L) \mid \\ & (V \mid R)(\bar{c}R \mid \bar{d})^*\bar{c}d(c \mid Ld)^*(V \mid L).\end{aligned}$$

Case 1. $w \in V \mid L \mid R$. Since $w \neq \varepsilon$, $w \in LR \mid L \mid R$. It follows that ww has one of $\bar{d}c, \bar{c}c, \bar{d}d$ as a factor, which contradicts $ww \in \mathrm{fac}(V)$ by Lemma 6. So this case is impossible.
Case 2. $w \in (V \mid R)(\bar{c}R \mid \bar{d})^*(\bar{c} \mid \bar{c}R \mid \bar{d})$. If w starts in c, then ww contains either $\bar{c}c$ or $\bar{d}c$ as a factor, which contradicts $ww \in \mathrm{fac}(V)$ by Lemma 6. So

$$w \in (\varepsilon \mid R)(\bar{c}R \mid \bar{d})^*(\bar{c} \mid \bar{c}R \mid \bar{d}).$$

Case 2.1. $w \in (\bar{c}R \mid \bar{d})^*(\bar{c} \mid \bar{c}R \mid \bar{d})$. If w ends in $\bar{c}$, ww contains either $\bar{c}\bar{c}$ or $\bar{c}\bar{d}$ as a factor, which contradicts $ww \in \mathrm{fac}(V)$ by Lemma 6. So in this case $w \in (\bar{c}R \mid \bar{d})^*(\bar{c}R \mid \bar{d}) = (\bar{c}R \mid \bar{d})^+$.
Case 2.2. $w \in R(\bar{c}R \mid \bar{d})^*(\bar{c} \mid \bar{c}R \mid \bar{d})$ In this case, w starts in d. If w ends in $\bar{d}$, then ww contains either $\bar{d}d$ as a factor, contradicting $ww \in \mathrm{fac}(V)$ by Lemma 6. So in this case $w \in R(\bar{c}R \mid \bar{d})^*\bar{c}$.

Case 3. $w \in (c \mid Ld \mid d)(c \mid Ld)^*(V \mid L)$. This case is exactly symmetric to Case 2, and we can conclude $w \in (c \mid Ld)^+$ or $w \in d(c \mid Ld)^*L$.

Case 4. $w \in (V \mid R)(\bar{c}R \mid \bar{d})^* \bar{c}d(c \mid Ld)^*(V \mid L)$. Let $m, n \geq 0$ be such that

$$w \in (V \mid R)(\bar{c}R \mid \bar{d})^m \bar{c}d(c \mid Ld)^n(V \mid L).$$

We show that $m \neq n$. Suppose, by way of contradiction, $m = n$. Then ww contains a factor u that belongs to

$$d(c \mid Ld)^n(V \mid L)(V \mid R)(\bar{c}R \mid \bar{d})^n \bar{c}.$$

Note that

$$u \rhd^* u' \in d(c \mid d)^n(\bar{c} \mid \bar{d})^n \bar{c}.$$

It is easy to see from this that $\mathrm{nf}(u)$ has either $c\bar{d}$ or $d\bar{c}$ as a factor. But since u is a factor of ww, $u \in \mathrm{fac}(V) \subseteq \mathrm{fac}(D_2^*)$. By Lemma 4, $\mathrm{nf}(u) \in (\bar{c} \mid \bar{d})^*(c \mid d)^*$, a contradiction.

We have proved that one of (i)–(v) holds in each case. □

3.2 Properties of the 3-MCFL H

Lemma 12 immediately yields a necessary condition for membership in $\{ w \in \hat{\Sigma}^+ \mid ww \in \mathrm{fac}(H) \}$. For w to be in this set, it must be that $\psi(w)\psi(w) = \psi(ww) \in \psi(\mathrm{fac}(H)) = \mathrm{fac}(\psi(H)) = \mathrm{fac}(V)$, so either $\psi(w) = \varepsilon$, in which case $w \in a^+ \mid b^+$, or $\psi(w)$ must satisfy one of the five conditions in Lemma 12. This will be used in the next section to give a necessary condition for membership in

$$\{ w \in \hat{\Sigma}^+ \mid ww \in \mathrm{fac}(H) \} \cap \mathrm{fac}(\{ v_n \mid n \in \mathbb{N} \}),$$

where $\{ v_n \mid n \in \mathbb{N} \}$ is a certain infinite subset of H. In this section, we establish some general properties of H that will be useful in the next section.

Lemma 13 *For every $v \in V$, there is a unique string $w \in H$ such that $\psi(w) = v$.*

Proof We prove by induction on the length of $v \in V$ that there is a unique triple (w_1, w_2, w_3) such that $J(w_1, w_2, w_3)$ is derivable and $\psi(w_2) = v$. It is clear from the grammar for H that $\vdash J(w_1, w_2, w_3)$ and $\psi(w_2) = \varepsilon$ imply $w_1 = a, w_2 = \varepsilon, w_3 = b$. This takes care of the case $v = \varepsilon$. Now suppose $v \in LR$. Then $v = cu_1\bar{c}du_2\bar{d}$ for some $u_1, u_2 \in V$. Note that the choice of u_1 and u_2 is unique. For, if $v = cu_1'\bar{c}du_2'\bar{d}$ for some $u_1', u_2' \in V$, then u_1' either is a prefix of u_1 or contains u_1 as a prefix, which implies $u_1 = u_1'$ by Lemma 10. Similarly, u_2' either is a suffix of u_2 or contains u_2 as a suffix, and it follows that $u_2 = u_2'$. If $\vdash J(w_1, w_2, w_3)$ and $\psi(w_2) = v$, then w_2 cannot be ε and there must be some $x_1, y_1 \in a^+$, $x_2, y_2 \in H$, and $x_3, y_3 \in b^+$ such that

$$\begin{aligned} &\vdash J(x_1, x_2, x_3), \\ &\vdash J(y_1, y_2, y_3), \\ w_1 &= ax_1, \\ w_2 &= y_1cx_2\bar{c}dy_2\bar{d}x_3, \\ w_3 &= y_3b. \end{aligned}$$

Since $\psi(w_2) = v$, we have $c\psi(x_2)\bar{c}d\psi(y_2)\bar{d} = cu_1\bar{c}du_2\bar{d}$. Since $x_2, y_2 \in H$, both $\psi(x_2)$ and $\psi(y_2)$ are in $\psi(H) = V$. It follows that $\psi(x_2) = u_1$ and $\psi(y_2) = u_2$. By induction hypothesis, (x_1,x_2,x_3) and (y_1,y_2,y_3) are uniquely determined by u_1 and u_2, respectively. Since u_1 and u_2 are uniquely determined by v, the triple (w_1,w_2,w_3) is uniquely determined by v. □

Let \$ be a symbol not in $\hat{\Sigma}$. We use this symbol to mark the beginning and end of a string in H.

Lemma 14 $\mathrm{fac}(\$H\$) \cap (\{\$\} \cup \hat{\Sigma})^2 = \{\$\$, \$a, aa, ac, b\$, bb, b\bar{c}, b\bar{d}, ca, c\bar{c}, \bar{c}d, da, d\bar{d}, \bar{d}b\}$.

Proof Let F denote the set on the right-hand side of the equation. We prove by induction on the length of u_2 that $\vdash J(u_1,u_2,u_3)$ implies $\mathrm{fac}(\$u_2\$) \cap (\{\$\} \cup \hat{\Sigma})^2 \subseteq F$. For the induction basis, observe that $\mathrm{fac}(\$\varepsilon\$) \cap (\{\$\} \cup \hat{\Sigma})^2 = \{\$\$\} \subseteq F$. Now suppose for some x_1,x_2,x_3,y_1,y_2,y_3 such that $\vdash J(x_1,x_2,x_3)$ and $\vdash J(y_1,y_2,y_3)$, we have $u_1 = ax_1, u_2 = y_1cx_2\bar{c}dy_2\bar{d}x_3, u_3 = y_3b$. It follows from the induction hypothesis applied to x_2 and y_2 that

$$\begin{aligned}
\mathrm{fac}(cx_2\bar{c}) \cap \hat{\Sigma}^2 &\subseteq (F - \{\$\$, \$a, b\$\}) \cup \{c\bar{c}, ca, b\bar{c}\} \\
&= F - \{\$\$, \$a, b\$\} \\
\mathrm{fac}(dy_2\bar{d}) \cap \hat{\Sigma}^2 &\subseteq (F - \{\$\$, \$a, b\$\}) \cup \{d\bar{d}, da, b\bar{d}\} \\
&= F - \{\$\$, \$a, b\$\}.
\end{aligned}$$

Since $y_1 \in a^+$ and $x_3 \in b^+$, we get

$$\begin{aligned}
&\mathrm{fac}(\$y_1cx_2\bar{c}dy_2\bar{d}x_3\$) \cap (\{\$\} \cup \hat{\Sigma})^2 \\
&\subseteq \{\$a, aa, ac\} \cup (\mathrm{fac}(cx_2\bar{c}) \cap \hat{\Sigma}^2) \cup \{\bar{c}d\} \cup (\mathrm{fac}(dy_2\bar{d}) \cap \hat{\Sigma}^2) \cup \{\bar{d}b, bb, b\$\} \\
&\subseteq F.
\end{aligned}$$

Therefore, $\mathrm{fac}(\$H\$) \cap (\{\$\} \cup \hat{\Sigma})^2 \subseteq F$. To see the converse inclusion, note that for $v = aacac\bar{c}d\bar{d}b\bar{c}dac\bar{c}d\bar{d}b\bar{d}bb \in H$, we have $\mathrm{fac}(\$v\$) \cap (\{\$\} \cup \hat{\Sigma})^2 = F - \{\$\$\}$. □

Lemma 15 *Let $u, w \in \hat{\Sigma}^*$ and $v \in \hat{\Sigma}^+$. If $uv \in H$ and $vw \in H$, then $u = w = \varepsilon$.*

Proof Since $v \neq \varepsilon$, Lemma 14 implies that both uv and vw start in a and end in b. Hence v starts in a and ends in b. By Lemma 14, the only symbols that can follow a in v are a and c, and the only symbols that can precede b in v are b and $\bar{d}$. So $v \in a^+c\hat{\Sigma}^*\bar{d}b^+$. Since $\psi(v) \neq \varepsilon$ and $\psi(uv)$ and $\psi(vw)$ are both in $\psi(H) = V$, Lemma 10 implies that $\psi(u) = \psi(w) = \varepsilon$. Hence $\psi(uv) = \psi(vw)$, and by Lemma 13, $uv = vw$. But $\psi(u) = \psi(w) = \varepsilon$ implies $u \in a^*$ and $w \in b^*$, and it easily follows that $u = w = \varepsilon$. □

Lemma 15 implies that H is both prefix-free and suffix-free.

Lemma 16 (i) $H \subseteq \varepsilon \mid (a^+c)^+\bar{c}\hat{\Sigma}^*d(\bar{d}b^+)^+$.
(ii) *If $\vdash J(u_1,u_2,u_3)$ and $u_2 \in (a^*c)^k(\bar{c}\hat{\Sigma}^*d \mid \varepsilon)(\bar{d}b^*)^l$, then $u_1 = a^{k+1}$ and $u_3 = b^{l+1}$.*

Proof (i). Suppose $v \neq \varepsilon$ and $v \in H$. We reason using Lemma 14. The first symbol of v must be a. Also, in v, the only symbols that can follow a are a and c, and the only symbols that can follow c are a and $\bar{c}$. Since the last symbol of v must be b, it follows that v has a prefix that belongs to $(a^+c)^+\bar{c}$. By a symmetric reasoning, v has a suffix that belongs to $d(\bar{d}b^+)^+$. Therefore, $v \in (a^+c)^+\bar{c}\hat{\Sigma}^*d(\bar{d}b^+)^+$.

(ii). We prove this part[7] by induction on the length of u_2. Suppose $\vdash J(u_1,u_2,u_3)$. If $u_2 = \varepsilon \in (a^*c)^0(\bar{c}\hat{\Sigma}^*d \mid \varepsilon)(\bar{d}b^*)^0$, then we must have $u_1 = a^1$ and $u_3 = b^1$. If $u \neq \varepsilon$, then there exist x_1,x_2,x_3,y_1,y_2,y_3 such that $\vdash J(x_1,x_2,x_3)$, $\vdash J(y_1,y_2,y_3)$, $u_1 = ax_1$, $u_2 = y_1cx_2\bar{c}dy_2\bar{d}x_3$, $u_3 = y_3b$. Suppose $u_2 \in (a^*c)^k(\bar{c}\hat{\Sigma}^*d \mid \varepsilon)(\bar{d}b^*)^l$. Since $y_1 \in a^*$ and $x_3 \in b^*$, we have $k,l \geq 1$, and part (i) of the lemma implies that for some $m,n \geq 0$,

$$x_2 \in (a^*c)^{k-1}(\bar{c}\hat{\Sigma}^*d \mid \varepsilon)(\bar{d}b^*)^m,$$
$$y_2 \in (a^*c)^n(\bar{c}\hat{\Sigma}^*d \mid \varepsilon)(\bar{d}b^*)^{l-1}.$$

By induction hypothesis, $x_1 = a^k$ and $y_3 = b^l$. Therefore, $u_1 = a^{k+1}$ and $u_3 = b^{l+1}$. □

Note that by Lemma 14, in any string in H, $\bar{c}$ always precedes d and d always follows $\bar{c}$.

Lemma 17 *For all $u,v \in \hat{\Sigma}^*$, the following conditions hold:*

(i) *If $ucv \in H$, then for some $k \geq 1$,*

$$u \in (\varepsilon \mid \hat{\Sigma}^*(c \mid d))a^k, \quad a^kcv \in H(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)$$

(ii) *If $u\bar{d}v \in H$, then for some $l \geq 1$,*

$$u\bar{d}b^l \in (\varepsilon \mid \hat{\Sigma}^*(c \mid d))H, \quad v \in b^l(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*).$$

(iii) *If $u\bar{c}dv \in H$, then for some $k,l \geq 1$,*

$$u \in (\varepsilon \mid \hat{\Sigma}^*(c \mid d))a^kcH, \quad v \in H\bar{d}b^l(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*).$$

Proof Each of the three conditions can be proved by easy induction on the combined length of u and v. We only prove (i). Suppose $ucv \in H$. Since $ucv \neq \varepsilon$, there must be $y_1 \in a^+$, $x_2,y_2 \in H$, and $x_3 \in b^+$ such that $ucv = y_1cx_2\bar{c}dy_2\bar{d}x_3$. If $u = y_1$, then we can take $a^k = y_1$. Otherwise, either $u = y_1cx_2'$, $v = x_2''\bar{c}dy_2\bar{d}x_3$ for some x_2', x_2'' such that $x_2 = x_2'cx_2''$, or $u = y_1cx_2\bar{c}dy_2'$, $v = y_2''\bar{d}x_3$ for some y_2', y_2'' such that $y_2 = y_2'cy_2''$. In the former case, we can apply the induction hypothesis to x_2', x_2'' and obtain $x_2' \in (\varepsilon \mid \hat{\Sigma}^*(c \mid d))a^k$ and $a^kcx_2'' \in H(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)$ for some $k \geq 1$. It follows that $u = y_1cx_2' \in \hat{\Sigma}^*(c \mid d)a^k$ and $a^kcv = a^kcx_2''\bar{c}dy_2\bar{d}x_3 \in H(\bar{c} \mid \bar{d})\hat{\Sigma}^*$. In the latter case, we can apply the induction hypothesis to y_2', y_2'' and obtain $y_2' \in (\varepsilon \mid \hat{\Sigma}^*(c \mid d))a^k$ and $a^kcy_2'' \in H(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)$ for some $k \geq 1$, and we can similarly infer $u = y_1cx_2\bar{c}dy_2' \in \hat{\Sigma}^*(c \mid d)a^k$ and $a^kcv = a^kcy_2''\bar{d}x_3 \in H(\bar{c} \mid \bar{d})\hat{\Sigma}^*$. □

Lemma 18 *Suppose $w \in \mathrm{fac}(\$H\$)$. For all $k,l \geq 0$, the following conditions hold:*

[7] By part (i), part (ii) can be equivalently stated with a^+ and b^+ in place of a^* and b^*, but it will turn out to be slightly more convenient in this form.

(i) $w \in (\$ \mid c \mid d)a^k cH\bar{c}d(a^*c)^l(\bar{c} \mid \bar{d})$ *implies* $k = l+1$.
(ii) $w \in (c \mid d)(\bar{d}b^*)^k\bar{c}dH\bar{d}b^l(\bar{c} \mid \bar{d} \mid \$)$ *implies* $k+1 = l$.

Proof We only prove part (i), since part (ii) is exactly symmetric. Suppose that $w \in \text{fac}(\$H\$)$ and for some $u \in H$,

$$w \in (\$ \mid c \mid d)w',$$
$$w' \in a^k cu\bar{c}d(a^*c)^l(\bar{c} \mid \bar{d}).$$

By Lemma 17, part (i), there is a string $z \in H$ such that w' is a prefix of some string in $z(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)$. Since w' starts in a or c, the string z cannot be ε. Hence there are some strings $x_1, x_2, x_3, y_1, y_2, y_3$ such that $\vdash J(x_1, x_2, x_3)$, $\vdash J(y_1, y_2, y_3)$, and $z = y_1cx_2\bar{c}dy_2\bar{d}x_3$. So

$$w' \text{ is a prefix of some string in } y_1cx_2\bar{c}dy_2\bar{d}x_3(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*).$$

Note that $x_1, y_1 \in a^+$ and $x_3, y_3 \in b^+$. So clearly, $y_1 = a^k$, and either $x_2\bar{c}$ is a prefix of $u\bar{c}$, or else $u\bar{c}$ is a prefix of $x_2\bar{c}$. Since $u \in H$ and $x_2 \in H$, neither u nor x_2 can start in $\bar{c}$. It follows that $u = \varepsilon$ if and only if $x_2 = \varepsilon$. If $u \neq \varepsilon$ and $x_2 \neq \varepsilon$, then either u is a non-empty prefix of x_2 or vice versa, and Lemma 15 implies that $u = x_2$. Hence we always have $a^k cu\bar{c}d = y_1cx_2\bar{c}d$. It follows that $y_2\bar{d}$ has a prefix belonging to $(a^*c)^l(\bar{c} \mid \bar{d})$. Since $y_2 \in H$, by Lemma 16, part (i), either $l = 0$ and $y_2 = \varepsilon$ or $l \geq 1$ and y_2 has a prefix belonging to $(a^*c)^l\bar{c}$. We can now apply Lemma 16, part (ii), to $J(y_1, y_2, y_3)$ and obtain $k = l+1$. □

3.3 Almost Anti-iterative Elements of H

Given a language K and a string $w \in K$, an *iteration tuple* for w in K is a tuple of strings $(u_0, w_1, u_1, \ldots, w_k, u_k)$ such that

- $w = u_0w_1u_1 \ldots w_ku_k$,
- $w_1 \ldots w_k \neq \varepsilon$, and
- $u_0w_1^iu_1 \ldots w_k^iu_k \in K$ for all $i \geq 0$.

The notion of an iteration tuple is a generalization of the notion of an *iterative pair* [1]. A language K is said to be *k-iterative* if all but finitely many strings in K have an iteration tuple $(u_0, w_1, u_1, \ldots, w_k, u_k)$ (of length $2k+1$) in K. We simply say that K is *iterative* if all but finitely many strings in K have an iteration tuple (of any length) in K. (Iterativity is a slight weakening of the property Groenink [5,4] called *finite pumpability*.)

We prove a theorem that implies that the language H is not iterative. In fact, the theorem states something much stronger. We say that a string $v \in K$ is *anti-iterative* in K if $v = u_0w_1u_1 \ldots w_ku_k$ and $w_1 \ldots w_k \neq \varepsilon$ (for any $k \geq 1$) imply $u_0w_1^iu_1 \ldots w_k^iu_k \notin K$ for all $i > 1$. We say that $v \in K$ is *almost anti-iterative* in K if $v = u_0w_1u_1 \ldots w_ku_k$ and $w_1 \ldots w_k \neq \varepsilon$ (for any $k \geq 1$) imply that there is at most one natural number $i > 1$ such that $u_0w_1^iu_1 \ldots w_k^iu_k \in K$. Clearly, if v is almost anti-iterative in K, then there is no iteration tuple for v in K.

Now for each $n \geq 0$, define a string $v_n \in H$ as follows:

$$v_0 = \varepsilon,$$
$$v_{n+1} = a^{n+1} c v_n \bar{c} d v_n \bar{d} b^{n+1}.$$

It is easy to see $\vdash J(a^{n+1}, v_n, b^{n+1})$ for all $n \in \mathbb{N}$. The strings v_n are precisely those elements of H that have a derivation tree whose immediate subtree is a perfect binary tree. We will show that each v_n is almost anti-iterative in H.

We start with some lemmas (Lemmas 19–22) stating some general properties of the strings v_n that are intuitively obvious from the way they are defined. We give a fairly rigorous proof to each of these lemmas.

Lemma 19 *$v_n \in (a^+c)^n(\bar{c}\hat{\Sigma}^* d \mid \varepsilon)(\bar{d}b^+)^n$ for all n.*

Proof For $n = 0$, $v_0 = \varepsilon = (a^+c)^0 \varepsilon (\bar{d}b^+)^0$, so the desired condition holds. For $n \geq 1$, we prove by induction on n that $v_n \in (a^+c)^n \bar{c}\hat{\Sigma}^* d(\bar{d}b^+)^n$. For $n = 1$, $v_1 = ac\bar{c}d\bar{d}b \in (a^+c)^1 \bar{c}\hat{\Sigma}^* d(\bar{d}b^+)^1$. For $n \geq 2$, assume $v_{n-1} \in (a^+c)^{n-1}\bar{c}\hat{\Sigma}^* d(\bar{d}b^+)^{n-1}$. Then $v_n = a^n c v_{n-1} \bar{c} d v_{n-1} \bar{d} b^n \in (a^+c)^n \bar{c}\hat{\Sigma}^* d(\bar{d}b^+)^n$. □

Lemma 20 $\mathrm{fac}(\{v_n \mid n \in \mathbb{N}\}) \cap H = \{v_n \mid n \in \mathbb{N}\}$.

Proof Clearly, it suffices to show the inclusion, $\mathrm{fac}(\{v_n \mid n \in \mathbb{N}\}) \cap H \subseteq \{v_n \mid n \in \mathbb{N}\}$. We prove by induction on $n \in \mathbb{N}$ that $w \in \mathrm{fac}(v_n) \cap H$ implies $w = v_k$ for some $k \leq n$. Since $v_0 = \varepsilon \in H$, the induction basis is immediate. Now assume $w \in H$ and w is a factor of $v_{n+1} = a^{n+1} c v_n \bar{c} d v_n \bar{d} b^{n+1}$. By Lemma 16, part (i), either $w = \varepsilon$ or $w \in (a^+c)^+ \bar{c}\hat{\Sigma}^* d(\bar{d}b^+)^+$. If $w = \varepsilon$, then $w = v_0$. It remains to consider the case where $w \in (a^+c)^+ \bar{c}\hat{\Sigma}^* d(\bar{d}b^+)^+$. If $\psi(w) = \psi(v_{n+1})$, then $w = v_{n+1}$ by Lemma 13. If $\psi(w) \neq \psi(v_{n+1})$, then either w is a factor of $v_n \bar{c} d v_n \bar{d} b^{n+1}$ or w is a factor of $a^{n+1} c v_n \bar{c} d v_n$.

Case 1. w is a factor of $v_n \bar{c} d v_n \bar{d} b^{n+1}$. Since w starts in a, there must be a non-empty suffix y of v_n starting in a such that w is a prefix of $y \bar{c} d v_n \bar{d} b^{n+1}$ or of $y \bar{d} b^{n+1}$. Since y is a suffix of $v_n \in H$, Lemma 15 implies that y cannot be a proper prefix of any element of H. Since $w \in H$, it follows that y is not a proper prefix of w. Since w is a prefix of $y \bar{c} d v_n \bar{d} b^{n+1}$ or of $y \bar{d} b^{n+1}$, w must be a prefix of y.
Case 2. w is a factor of $a^{n+1} c v_n \bar{c} d v_n$. Since w ends in b, there must be a non-empty prefix x of v_n ending in b such that w is a suffix of $a^{n+1} c x$ or of $a^{n+1} c v_n \bar{c} d x$. By an analogous reasoning to the previous case, we can conclude that w is a suffix of x.

In both cases, w is a factor of v_n, and the induction hypothesis gives $w = v_k$ for some $k \leq n$. □

Lemma 21 *Suppose $w \in \mathrm{fac}(\$\{v_n \mid n \in \mathbb{N}\}\$)$. For all $k, l \geq 0$, the following conditions hold:*

(i) *$w \in (\$ \mid c \mid d) a^k (c \mid cH\bar{c}d) a^l (c \mid \bar{c} \mid \bar{d})$ implies $k = l + 1$.*
(ii) *$w \in (c \mid d \mid \bar{d}) b^k (\bar{c}dH\bar{d} \mid \bar{d}) b^l (\bar{c} \mid \bar{d} \mid \$)$ implies $k + 1 = l$.*
(iii) *$w \in (c \mid \bar{d}) b^k \bar{c} d a^l (c \mid \bar{d})$ implies $k = l$.*

Proof (i). Suppose $uwv = \$v_n\$$ and

$$\begin{aligned} w &\in (\$ \mid c \mid d)w', \\ w' &\in a^k(c \mid cH\bar{c}d)a^l(c \mid \bar{c} \mid \bar{d}). \end{aligned} \tag{10}$$

By Lemma 17, part (i), $k \geq 1$ and there is a $z \in H$ such that $w'v \in z(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)\$$. Since w' starts in a, $z \neq \varepsilon$. Lemma 20 implies that $z = v_k = a^k c v_{k-1} \bar{c} d v_{k-1} \bar{d} b^k$. So

$$w'v \in a^k c v_{k-1} \bar{c} d v_{k-1} \bar{d} b^k (\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)\$. \tag{11}$$

By (10), either $w' \in a^k c a^l (c \mid \bar{c} \mid \bar{d})$ or $w' \in a^k cH\bar{c}da^l(c \mid \bar{c} \mid \bar{d})$.

Case 1. $w' \in a^k c a^l (c \mid \bar{c} \mid \bar{d})$. Then either $k = 1$, $v_{k-1} = \varepsilon$, $l = 0$, and $w' = a^k c\bar{c}$, or $k \geq 2$ and v_{k-1} has a prefix that belongs to $a^l(c \mid \bar{c} \mid d)$, which implies $l = k-1$. In either case, we get $k = l+1$.
Case 2. $w' \in a^k c x \bar{c} d a^l (c \mid \bar{c} \mid \bar{d})$ for some $x \in H$. Then either $v_{k-1}\bar{c}$ is a prefix of $x\bar{c}$ or $x\bar{c}$ is a prefix of $v_{k-1}\bar{c}$. Since neither v_{k-1} nor x can start in $\bar{c}$, it follows that $v_{k-1} = \varepsilon$ if and only if $x = \varepsilon$. If $v_{k-1} \neq \varepsilon$ and $x \neq \varepsilon$, then either v_{k-1} is a non-empty prefix of x or x is a non-empty prefix of v_{k-1}. Lemma 15 then implies $v_{k-1} = x$. So we always have $a^k c v_{k-1} \bar{c} d = a^k c x \bar{c} d$. By (11), it follows that $v_{k-1}\bar{d}$ has a prefix that belongs to $a^l(c \mid \bar{c} \mid \bar{d})$. But the definition of v_n implies that $v_{k-1}\bar{d}$ always has a prefix in $a^{k-1}(c \mid \bar{d})$. Therefore, $l = k-1$ and so $k = l+1$.

(ii). Exactly symmetric to part (i).
(iii). Suppose $uwv = \$v_n\$$ and

$$\begin{gathered} w = w'\bar{c}dw'', \\ w' \in (c \mid \bar{d})b^k, \qquad w'' \in a^l(c \mid \bar{d}). \end{gathered}$$

By Lemma 17, part (iii), there exist $x, y \in H$ and $k', l' \geq 1$ such that

$$uw' \in \$(\varepsilon \mid \hat{\Sigma}^*(c \mid d))a^{k'}cx, \qquad w''v \in y\bar{d}b^{l'}(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)\$.$$

Since x and y are factors of v_n, Lemma 20 implies that $x = v_i$ and $y = v_j$ for some $i, j \geq 0$. If $i \geq 1$, then v_i has $\bar{d}b^i$ as a suffix, so it follows that $k = i$. If $i = 0$, then uw' ends in c, so $w' = c$ and $k = 0$. So we always have $k = i$. By a symmetric reasoning, we get $l = j$. It follows that

$$uwv = uw'\bar{c}dw''v \in \$(\varepsilon \mid \hat{\Sigma}^*(c \mid d))a^{k'}cv_k\bar{c}dv_l\bar{d}b^{l'}(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)\$.$$

Since $v_k\bar{c}$ has a prefix that belongs to $a^k(c \mid \bar{c})$ and $v_l\bar{d}$ has a prefix that belongs to $a^l(c \mid \bar{d})$, part (i) of this lemma implies $k' = k+1 = l+1$. Therefore, $k = l$. □

We will make frequent use of Lemmas 18 and 21 in what follows. It will be important not to confuse part (i) and (ii) of Lemma 18, on the one hand, and part (i) and (ii) of Lemma 21, on the other. The former state general properties of elements of H, while the latter express special properties of the strings v_n.

Lemma 22 *Suppose* $w \in \mathrm{fac}(\{\, v_n \mid n \in \mathbb{N}\,\})$.

(i) *If* $\psi(w) \in L$*, then* $w = a^i cv_k\bar{c}$ *for some* $i,k \geq 0$ *such that* $i \leq k+1$.
(ii) *If* $\psi(w) \in R$*, then* $w = dv_k\bar{d}b^j$ *for some* $j,k \geq 0$ *such that* $j \leq k+1$.
(iii) *If* $\psi(w) \in LR$*, then* $w = a^i cv_k\bar{c}dv_k\bar{d}b^j$ *for some* $i,j,k \geq 0$ *such that* $i,j \leq k+1$.

Proof (i). Suppose $uwv = v_n$ and $\psi(w) \in L = cV\bar{c}$. By Lemma 14, in the string w, b cannot precede a or c and neither a nor b can follow $\bar{c}$. Hence $w = a^i cx\bar{c}$ for some $i \in \mathbb{N}$ and some x such that $\psi(x) \in V$.

Since $uwv = ua^i cx\bar{c}v = v_n \in H$, Lemma 17, part (i), implies that there must be some $l \geq 1$ and $y \in H$ such that $l \geq i$, a^l is a suffix of ua^i and $a^l cx\bar{c}v \in y(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*)$. This means that y must contain $a^l c$ as a prefix, so Lemma 20 implies $y = v_l = a^l cv_{l-1}\bar{c}dv_{l-1}\bar{d}b^l$. Hence

$$a^l cx\bar{c}v \in a^l cv_{l-1}\bar{c}dv_{l-1}\bar{d}b^l(\varepsilon \mid (\bar{c} \mid \bar{d})\hat{\Sigma}^*).$$

This implies the following:

$$\text{Either } x\bar{c} \text{ is a prefix of } v_{l-1}\bar{c}\text{, or else } v_{l-1}\bar{c} \text{ is a prefix of } x\bar{c}. \tag{12}$$

We claim $x = v_{l-1}$. The desired conclusion follows from this by putting $k = l-1$.

Case 1. $l = 1$. Then $v_{l-1} = v_0 = \varepsilon$. Since $\psi(x) \in V$ implies that x cannot start in $\bar{c}$, it is clear from (12) that x must be ε. So the claim holds in this case.
Case 2. $l \geq 2$. It follows from (12) that either $\psi(x)\bar{c}$ is a prefix of $\psi(v_{l-1})\bar{c}$ or vice versa. Since $l-1 \geq 1$, $\psi(v_{l-1})$ starts in c. Then $\psi(x)$ must also start in c. Hence either $\psi(x)$ is a non-empty prefix of $\psi(v_{l-1})$ or $\psi(v_{l-1})$ is a non-empty prefix of $\psi(x)$. By Lemma 10, we get $\psi(v_{l-1}) = \psi(x)$. Consequently, $x\bar{c}$ is not a prefix of v_{l-1}, and $v_{l-1}\bar{c}$ is not a prefix of x, so by (12), we can conclude $v_{l-1} = x$.

(ii). This is proved in an exactly symmetric way to (i).
(iii). By Part (i) and (ii) of this lemma, $w = a^i cv_k\bar{c}dv_l\bar{d}b^j$ for some $i,j,k \geq 0$ such that $i \leq k+1$ and $j \leq l+1$. Since w contains a factor that belongs to $(c \mid \bar{d})b^k\bar{c}da^l(c \mid \bar{d})$, part (iii) of Lemma 21 gives $k = l$. □

We now state and prove our main lemma. Let

$$\widehat{L} = \{\, cv_n\bar{c} \mid n \in \mathbb{N} \,\},$$
$$\widehat{R} = \{\, dv_n\bar{d} \mid n \in \mathbb{N} \,\},$$
$$\widehat{LR} = \{\, cv_n\bar{c}dv_n\bar{d} \mid n \in \mathbb{N} \,\}.$$

Then Lemma 22 implies

$$\psi^{-1}(L) \cap \mathrm{fac}(\{\, v_n \mid n \in \mathbb{N} \,\}) \subseteq a^*\widehat{L}, \tag{13}$$
$$\psi^{-1}(R) \cap \mathrm{fac}(\{\, v_n \mid n \in \mathbb{N} \,\}) \subseteq \widehat{R}b^*, \tag{14}$$
$$\psi^{-1}(LR) \cap \mathrm{fac}(\{\, v_n \mid n \in \mathbb{N} \,\}) \subseteq a^*\widehat{LR}b^*. \tag{15}$$

Lemma 23 *If* $w \in \mathrm{fac}(\{\, v_n \mid n \in \mathbb{N} \,\})$ *and* $ww \in \mathrm{fac}(H)$*, then*

$$\psi(w) \in c^* \mid Ldc^* \mid \bar{d}^* \mid \bar{d}^*\bar{c}R \mid V\bar{c}dc^+ \mid \bar{d}^+\bar{c}dV.$$

Proof Since ε clearly belongs to the required set, assume $\psi(w) \in \Sigma^+$. Since $ww \in \text{fac}(H)$ implies $\psi(w)\psi(w) \in \text{fac}(V)$, $\psi(w)$ must satisfy one of the five cases of Lemma 12:

1. $\psi(w) \in (\bar{c}R \mid \bar{d})^+$.
2. $\psi(w) \in R(\bar{c}R \mid \bar{d})^*\bar{c}$.
3. $\psi(w) \in (c \mid Ld)^+$.
4. $\psi(w) \in d(c \mid Ld)^*L$.
5. $\psi(w) \in (V \mid R)(\bar{c}R \mid \bar{d})^m\bar{c}d(c \mid Ld)^n(V \mid L)$ for some $m, n \geq 0$ such that $m \neq n$.

Below we treat the five cases in turn.

Case 1. $\psi(w) \in (\bar{c}R \mid \bar{d})^+$. We show that $\psi(w) \in \bar{d}^+ \mid \bar{d}^*\bar{c}R$. Suppose by way of contradiction that $\psi(w) \in \bar{d}^*\bar{c}R(\bar{c}R \mid \bar{d})^+$. Lemma 14 says that in the string w, a cannot precede $\bar{d}$ or $\bar{c}$, b can follow only $\bar{d}$, and $\bar{d}$ can be followed only by b. Together with (14), this allows us to infer

$$w \in b^*(\bar{d}b^+)^*\bar{c}\widehat{R}b^+((\bar{c}\widehat{R} \mid \bar{d})b^+)^*(\bar{c}\widehat{R} \mid \bar{d})b^*.$$

Recall that $\widehat{R}$ consists of the strings $dv_i\bar{d}$. Recall also that $v_i = \varepsilon$ when $i = 0$ and $v_i = a^i c v_{i-1}\bar{c}dv_{i-1}\bar{d}b^i$ otherwise. So if w contains a factor that belongs to

$$dv_i\bar{d}b^j(\bar{c} \mid \bar{d}),$$

then w contains a factor that belongs to

$$(d \mid \bar{d})b^i\bar{d}b^j(\bar{c} \mid \bar{d}),$$

and part (ii) of Lemma 21 allows us to infer $j = i + 1$. Hence w must be of the form[8]

$$w = ux_1 \dots x_m\bar{c}dv_k\bar{d}b^{k+1}y_1 \dots y_n z,$$

where $m, n \geq 0$ and

$$\begin{aligned} u &\in b^*, \\ x_i &= \bar{d}b^{p_i} \quad \text{for some } p_i \geq 1, \\ y_i &\in (\bar{c}dv_{q_i}\bar{d} \mid \bar{d})b^{q_i+1} \quad \text{for some } q_i \geq 0, \\ z &\in (\bar{c}dv_l\bar{d} \mid \bar{d})b^* \quad \text{for some } l \geq 0. \end{aligned}$$

Lemma 21, part (ii), also implies

$$\begin{aligned} q_{i+1} &= q_i + 1 \quad \text{for } i = 1, \dots, n-1, \\ q_1 &= k + 1 \quad \text{if } n \geq 1. \end{aligned}$$

So

$$q_i = k + i \quad \text{for } i = 1, \dots, n.$$

[8] We will appeal to Lemma 21 similarly in Cases 2–5 without explicitly going through this kind of reasoning.

It immediately follows that

$$\bar{d}b^{k+1}y_1 \dots y_n \text{ contains } \bar{d}b^{k+n+1} \text{ as a suffix.} \tag{16}$$

Note that this holds even when $n = 0$.

Next, we claim that

$$dv_k\bar{d}b^{k+1}y_1 \dots y_n \text{ has a suffix that belongs to } d(\bar{d}b^*)^{k+n+1}. \tag{17}$$

By Lemma 19, this is clearly true when $n = 0$. When $n \geq 1$, we can prove by induction on $i \in \{1, \dots, n\}$ that $dv_k\bar{d}b^{k+1}y_1 \dots y_i$ always has a suffix in $d(\bar{d}b^*)^{k+i+1}$. For $i = 0$, $dv_k\bar{d}b^{k+1}$ has a suffix in $d(\bar{d}b^*)^{k+1}$ by Lemma 19. For $1 \leq i \leq n$, assume that $dv_k\bar{d}b^{k+1}y_1 \dots y_{i-1}$ has a suffix in $d(\bar{d}b^*)^{k+i}$. If $y_i = \bar{d}b^{q_i+1} = \bar{d}b^{k+i}$, then it follows that $dv_k\bar{d}b^{k+1}y_1 \dots y_i$ has a suffix in $d(\bar{d}b^*)^{k+i+1}$. If $y_i = \bar{c}dv_{q_i}\bar{d}b^{q_i+1} = \bar{c}dv_{k+i}\bar{d}b^{k+i+1}$, then y_i has a suffix in $d(\bar{d}b^*)^{k+i+1}$ by Lemma 19.

Now note that

$$ww \text{ has a factor in } \bar{c}dv_k\bar{d}b^{k+1}y_1 \dots y_n zux_1 \dots x_m\bar{c}dv_k\bar{d}b^{k+1}(\bar{c} \mid \bar{d}). \tag{18}$$

Since $ww \in \text{fac}(H)$, this factor must also belong to $\text{fac}(H)$. We distinguish two cases.

Case 1.1. $z \in \bar{c}dv_l\bar{d}b^*$. Then by Lemma 19, $zux_1 \dots x_m$ has a suffix in $d(\bar{d}b^*)^{l+1+m}$, so by Lemma 18, part (ii), we get $l+1+m+1 = k+1$, i.e.,

$$k = l+m+1. \tag{19}$$

By (16), w contains as a factor

$$\bar{d}b^{k+n+1}z \in \bar{d}b^{k+n+1}\bar{c}dv_l\bar{d}b^*.$$

Since this factor belongs to $\text{fac}(\{\, v_n \mid n \in \mathbb{N}\,\})$, we must have

$$l = k+n+1$$

by Lemma 21, part (iii). But this last equation contradicts (19).

Case 1.2. $z \in \bar{d}b^*$. By (17), we see that $dv_k\bar{d}b^{k+1}y_1 \dots y_n zux_1 \dots x_m$ has a suffix in $d(\bar{d}b^*)^{k+n+1+1+m} = d(\bar{d}b^*)^{k+n+m+2}$. By Lemma 18, part (ii), we obtain from (18) that $k+n+m+2+1 = k+1$, a contradiction.

We have derived a contradiction in each case. So the assumption that $\psi(w) \in \bar{d}^*\bar{c}R(\bar{c}R \mid \bar{d})^+$ is incorrect and $\psi(w)$ must be in $\bar{d}^+ \mid \bar{d}^*\bar{c}R$.

Case 2. $\psi(w) \in R(\bar{c}R \mid \bar{d})^*\bar{c}$. We derive a contradiction. By Lemma 14, in the string w, $\bar{c}$ can be followed only by d and $\bar{d}$ can be followed only by b. Together with (14), this allows us to infer

$$w \in \widehat{R}b^+((\bar{c}\widehat{R} \mid \bar{d})b^+)^*\bar{c}.$$

By Lemma 21, part (ii), w must be of the form

$$w = dv_k\bar{d}b^{k+1}y_1 \dots y_n\bar{c},$$

where $n \geq 0$ and

$$y_i \in (\bar{c}dv_{q_i}\bar{d} \mid \bar{d})b^{q_i+1} \quad \text{for some } q_i \geq 0.$$

Lemma 21, part (ii), also implies

$$q_{i+1} = q_i + 1 \quad \text{for } i = 1, \dots, n-1,$$
$$q_1 = k+1 \quad \text{if } n \geq 1.$$

So we have

$$q_i = k+i \quad \text{for } i = 1, \dots, n.$$

As in Case 1, we can see that $v_k\bar{d}b^{k+1}y_1 \dots y_n$ has a suffix that belongs to $d(\bar{d}b^*)^{k+n+1}$. Since ww has a factor in

$$v_k\bar{d}b^{k+1}y_1 \dots y_n\bar{c}dv_k\bar{d}b^{k+1}(\bar{c} \mid \bar{d})$$

and this factor belongs to $\mathrm{fac}(H)$, Lemma 18, part (ii), implies $k+n+1+1 = k+1$, a contradiction.

Case 3. $\psi(w) \in (c \mid Ld)^+$. This case is exactly symmetric to Case 1 and we can derive $\psi(w) \in c^+ \mid Ldc^*$.

Case 4. $\psi(w) \in d(c \mid Ld)^*L$. This case is exactly symmetric to Case 2 and we can derive a contradiction.

Case 5. $\psi(w) \in (V \mid R)(\bar{c}R \mid \bar{d})^m\bar{c}d(c \mid Ld)^n(V \mid L)$ for some $m, n \geq 0$ such that $m \neq n$. We show that $\psi(w) \in \bar{d}^+\bar{c}dV \mid V\bar{c}dc^+$. By Lemma 14, a cannot precede $\bar{c}$ or $\bar{d}$, and b cannot follow c or d. Together with (13), (14), and (15), this allows us to infer

$$w \in (b^* \mid a^*\widehat{LR}b^* \mid \widehat{R}b^*)((\bar{c}\widehat{R} \mid \bar{d})b^*)^m\bar{c}d(a^*(c \mid \widehat{L}d))^n(a^* \mid a^*\widehat{LR}b^* \mid a^*\widehat{L}).$$

By Lemma 21, part (i) and (ii), we can write w as

$$w = xx_1 \dots x_m\bar{c}dy_n \dots y_1y,$$

where

$$x \in b^* \mid a^*cv_k\bar{c}dv_k\bar{d}b^{k+1} \mid dv_k\bar{d}b^{k+1} \quad \text{for some } k \geq 0,$$
$$y \in a^* \mid a^{l+1}cv_l\bar{c}dv_l\bar{d}b^* \mid a^{l+1}cv_l\bar{c} \quad \text{for some } l \geq 0,$$
$$x_i \in (\bar{c}dv_{p_i}\bar{d} \mid \bar{d})b^{p_i+1} \quad \text{for some } p_i \geq 0,$$
$$y_i \in a^{q_i+1}(c \mid cv_{q_i}\bar{c}d) \quad \text{for some } q_i \geq 0.$$

Lemma 21, part (i) and (ii), also implies

$$p_{i+1} = p_i + 1 \quad \text{for } i = 1, \dots, m-1, \tag{20}$$
$$q_{i+1} = q_i + 1 \quad \text{for } i = 1, \dots, n-1. \tag{21}$$

We first show that

$$yx = v_j \quad \text{for some } j. \tag{22}$$

Since ww contains $dy_n \dots y_1 yxx_1 \dots x_m \bar{c}$ as a factor and $ww \in \text{fac}(H)$,

$$(c \mid d)yx(\bar{c} \mid \bar{d}) \cap \text{fac}(H) \neq \varnothing. \tag{23}$$

By Lemma 14, the only symbol that can follow $\bar{c}$ in yx is d and the only symbol that can precede d in yx is $\bar{c}$. So $x = dv_k\bar{d}b^{k+1}$ if and only if $y = a^{l+1}cv_l\bar{c}$. Lemma 14 also implies that neither a nor c can follow b or $\bar{d}$ in yx, so we cannot have both $x \in a^*cv_k\bar{c}dv_k\bar{d}b^{k+1}$ and $y \in a^{l+1}cv_l\bar{c}dv_l\bar{d}b^*$. Hence

$$yx \in a^*b^* \mid a^*cv_k\bar{c}dv_k\bar{d}b^{k+1} \mid a^{l+1}cv_l\bar{c}dv_l\bar{d}b^* \mid a^{l+1}cv_l\bar{c}dv_k\bar{d}b^{k+1}.$$

If $yx \in a^*b^*$, Lemma 14 together with (23) implies $yx = \varepsilon = v_0$. Otherwise, Lemmas 18 and 19 together with (23) imply

$$yx = a^{j+1}cv_j\bar{c}dv_j\bar{d}b^{j+1} = v_{j+1},$$

where $j = k$ or $j = l$. This establishes (22).

Since $m \neq n$, either $m \geq 1$ or $n \geq 1$. We distinguish three cases:

Case 5.1. $m \geq 1, n \geq 1$. In this case, ww contains a factor in

$$(c \mid d)y_1v_jx_1(\bar{c} \mid \bar{d}).$$

This factor is in $\text{fac}(H)$. Since $\psi(ww) \in \text{fac}(V) \subseteq \text{fac}(D_2^*)$, we have $\psi(y_1v_jx_1) \in \text{fac}(D_2^*)$. By Lemma 4, $\text{nf}(\psi(y_1v_jx_1)) \in (\bar{c} \mid \bar{d})^*(c \mid d)^*$, and it follows that

$$y_1v_jx_1 \in a^{q_1+1}cv_j\bar{c}dv_{p_1}\bar{d}b^{p_1+1} \mid a^{q_1+1}cv_{q_1}\bar{c}dv_j\bar{d}b^{p_1+1}.$$

So

$$(c \mid d)(a^{q_1+1}cv_j\bar{c}dv_{p_1}\bar{d}b^{p_1+1} \mid a^{q_1+1}cv_{q_1}\bar{c}dv_j\bar{d}b^{p_1+1})(\bar{c} \mid \bar{d}) \cap \text{fac}(H) \neq \varnothing.$$

By Lemmas 18 and 19, we obtain $p_1 = q_1 = j$. By (20) and (21), then, we get $p_m = j + m - 1$ and $q_n = j + n - 1$. Since

$$x_m\bar{c}dy_n \in (\bar{c}dv_{j+m-1}\bar{d} \mid \bar{d})b^{j+m}\bar{c}da^{j+n}(c \mid cv_{j+n-1}\bar{c}d)$$

is a factor of w, we get $j + m = j + n$ by Lemma 21, part (iii), but this contradicts $m \neq n$.

Case 5.2. $m \geq 1, n = 0$. Since

$$ww = xx_1 \ldots x_m \bar{c} d v_j x_1 \ldots x_m \bar{c} d y$$

and $\psi(ww) \in \text{fac}(V) \subseteq \text{fac}(D_2^*)$, we get $\psi(dv_j x_1) \in \text{fac}(D_2^*)$. By Lemma 4, $\text{nf}(\psi(dv_j x_1)) = \text{nf}(d\psi(x_1)) \in (\bar{c} \mid \bar{d})^*(c \mid d)^*$. Hence we must have

$$x_1 = \bar{d} b^{p_1+1}.$$

By (20), $p_i = p_1 + i - 1$ for $i = 1, \ldots, m$. We consider three subcases, depending on whether $x \in b^*$, and whether $x_i = \bar{d} b^{p_1+i}$ for all $i = 1, \ldots, m$.
Case 5.2.1. $x \in b^*$ and $x_i = \bar{d} b^{p_1+i}$ for all $i = 1, \ldots, m$. Then since $yx = v_j$, either $x = y = \varepsilon$ or $j = l + 1$ and $y \in a^{l+1} c v_l \bar{c} d v_l \bar{d} b^*$. Hence

$$\psi(w) \in \bar{d}^+ \bar{c} dV.$$

Case 5.2.2. $x \notin b^*$ and $x_i = \bar{d} b^{p_1+i}$ for all $i = 1, \ldots, m$. Then $j = k + 1$, $yx = v_{k+1}$, and $dv_k \bar{d} b^{k+1}$ is a suffix of x. Since w contains a factor in

$$dv_k \bar{d} b^{k+1} x_1 (\bar{c} \mid \bar{d}) = dv_k \bar{d} b^{k+1} \bar{d} b^{p_1+1} (\bar{c} \mid \bar{d}),$$

we get $p_1 = k + 1$ by Lemma 21, part (ii). By Lemma 19, we also see that $xx_1 \ldots x_m$ has a suffix in $d(\bar{d} b^*)^{k+m+1}$. Since ww has a factor in

$$\begin{aligned} xx_1 \ldots x_m \bar{c} d v_{k+1} x_1 (\bar{c} \mid \bar{d}) &= xx_1 \ldots x_m \bar{c} d v_{k+1} \bar{d} b^{k+2} (\bar{c} \mid \bar{d}) \\ &\subseteq \hat{\Sigma}^* d(\bar{d} b^*)^{k+m+1} \bar{c} d H \bar{d} b^{k+2} (\bar{c} \mid \bar{d}), \end{aligned}$$

we get by Lemma 18, part (ii),

$$k + m + 1 + 1 = k + 2,$$

which contradicts $m \geq 1$.
Case 5.2.3. $x_h = \bar{c} d v_{p_1+h-1} \bar{d} b^{p_1+h}$ for some $h \in \{2, \ldots, m\}$. (Recall $x_1 = \bar{d} b^{p_1+1}$.) We can assume h to be the largest such number, i.e., $x_i = \bar{d} b^{p_1+i}$ for all $i \in \{h+1, \ldots, m\}$. By Lemma 19, x_h has a suffix in $d(\bar{d} b^*)^{p_1+h}$. It follows that $x_h \ldots x_m$ has a suffix in $d(\bar{d} b^*)^{p_1+m}$. Since ww has a factor in

$$\begin{aligned} x_h \ldots x_m \bar{c} d v_j x_1 (\bar{c} \mid \bar{d}) &= x_h \ldots x_m \bar{c} d v_j \bar{d} b^{p_1+1} (\bar{c} \mid \bar{d}) \\ &\subseteq \hat{\Sigma}^* d(\bar{d} b^*)^{p_1+m} \bar{c} d H \bar{d} b^{p_1+1} (\bar{c} \mid \bar{d}), \end{aligned}$$

we get by Lemma 18, part (ii),

$$p_1 + m + 1 = p_1 + 1,$$

which contradicts $m \geq 1$.
Case 5.3. $m = 0, n \geq 1$. This case is exactly symmetric to the preceding case, and we can conclude

$$\psi(w) \in V \bar{c} d c^+.$$

This concludes the proof of the lemma. □

Theorem 24 *For each $n \geq 0$, the string v_n is almost anti-iterative in H.*

Before embarking on the proof of the theorem, let us consider a simple example:

$$v_2 = \underbrace{aac}_{w_1}\,\underbrace{ac\bar{c}d\bar{d}}_{u_1}\,\underbrace{b\bar{c}dac\bar{c}d\bar{d}b\bar{d}b}_{w_2}\,\underbrace{b}_{w_3}.$$

In this example, $u_0 = u_2 = u_3 = \varepsilon$. Note

$$\psi(w_1) = c, \quad \psi(w_2) \in \bar{c}R, \quad \psi(w_3) = \varepsilon.$$

We have

$$w_1^2u_1w_2^2w_3^2 = aac\;\underbrace{aac\;\underbrace{ac\bar{c}d\bar{d}\,b}_{v_1}\bar{c}d\underbrace{ac\bar{c}d\bar{d}b}_{v_1}\bar{d}b\;b}_{v_2}\bar{c}d\underbrace{ac\bar{c}d\bar{d}b}_{v_1}\bar{d}b\;b\;b\;\in H,$$

but

$$w_1^3u_1w_2^3w_3^3 =$$
$$aac\;\underbrace{aac\;\underbrace{aac\;\underbrace{ac\bar{c}d\bar{d}\,b}_{v_1}\bar{c}d\underbrace{ac\bar{c}d\bar{d}b}_{v_1}\bar{d}b\;b}_{v_2}\bar{c}d\underbrace{ac\bar{c}d\bar{d}b}_{v_1}\bar{d}b\;b}_{\notin H}\bar{c}d\underbrace{ac\bar{c}d\bar{d}b}_{v_1}\bar{d}b\;b\;b\;b\;\notin H$$

After the occurrence of $\bar{d}$ following the third occurrence of v_1, one should find b^3, rather than b^2, in order to have a string in H (as required by Lemma 18, part (ii)).

Proof (of Theorem 24) Suppose that $v_n = u_0w_1u_1\dots w_ku_k$ and $w_1\dots w_k \neq \varepsilon$. If there is some j such that w_j^3 is not in $\mathrm{fac}(H)$, then there is no $i \geq 3$ such that $u_0w_1^iu_1\dots w_k^iu_k \in H$, and the conclusion of the theorem is clearly satisfied. Hence we may assume that each w_j^3 belongs to $\mathrm{fac}(H)$.

Suppose that $u_0w_1^h\dots w_k^hu_k \in H$ for some $h > 1$. We show that such h is unique.

Since w_j^2 is a factor of w_j^3 and hence belongs to $\mathrm{fac}(H)$, by Lemma 23, each $\psi(w_j)$ must belong to one of the six sets

$$c^*, \qquad Ldc^*, \qquad \bar{d}^*, \qquad \bar{d}^*\bar{c}R, \qquad V\bar{c}dc^+, \qquad \bar{d}^+\bar{c}dV.$$

Since $w_1\dots w_k \neq \varepsilon$, we have $u_0w_1u_1\dots w_ku_k \neq u_0w_1^hu_1\dots w_k^hu_k$. By Lemma 13, we know that $\psi(u_0w_1u_1\dots w_ku_k) \neq \psi(u_0w_1^hu_1\dots w_k^hu_k)$. Therefore, it cannot be that $\psi(w_j) = \varepsilon$ for all j. Since both $\psi(u_0w_1u_1\dots w_ku_k)$ and $\psi(u_0w_1^hu_1\dots w_k^hu_k)$ belong to V, the string $\psi(w_1)\dots\psi(w_k)$ must have the same number of occurrences of $c, \bar{c}, d, \bar{d}$. It follows that there is a j such that $\psi(w_j) \in Ldc^* \mid \bar{d}^*\bar{c}R \mid V\bar{c}dc^+ \mid \bar{d}^+\bar{c}dV$.

Case 1. $\psi(w_j) \in Ldc^*$. Lemma 14 implies that in the string w_j, b can follow only $\bar{d}$. So

$$w_j \in vd(a^*c)^*a^*$$

for some $v \in \mathrm{fac}(\{ v_n \mid n \in \mathbb{N} \})$ such that $\psi(v) \in L$. By Lemma 22, $v \in a^*cv_l\bar{c}$ for some $l \geq 0$. Lemma 14 also implies that in $u_0w_1u_1 \ldots w_ku_k$, (i) the only symbols that can precede a are a, c, and d, (ii) the only symbols that can follow a are a and c, and (iii) the only symbols that can follow c or d are a, $\bar{c}$, and $\bar{d}$. Hence we can write

$$\begin{aligned} u_0w_1u_1 \ldots w_{j-1}u_{j-1} &\in (\varepsilon \mid \hat{\Sigma}^*(c \mid d))a^{m_0}, \\ w_j &\in a^{m_1}cv_l\bar{c}d(a^*c)^p a^{m_2}, \\ u_jw_{j+1}u_{j+1} \ldots w_ku_k &\in (a^*c)^q(\bar{c} \mid \bar{d})\hat{\Sigma}^*, \end{aligned}$$

for some $l, m_0, m_1, m_2, p, q \geq 0$. We get $m_0 + m_1 = l + 1$ by Lemma 21, part (i), and $m_0 + m_1 = p + q + 1$ by Lemma 18, part (i). Hence $l = p + q$.

Let $g \geq j$ the largest number such that $u_jw_{j+1} \ldots u_{g-1}w_g \in (a^*c)^*a^*$. Let r be the number of occurrences of c in $w_{j+1} \ldots w_g$. Then for every $i \geq 1$,

$$u_jw_{j+1}^iu_{j+1} \ldots w_k^iu_k \in (a^*c)^{q+(i-1)r}(\bar{c} \mid \bar{d})\hat{\Sigma}^*.$$

Thus, $w_j^hu_jw_{j+1}^hu_{j+1} \ldots w_k^hu_k$ has a factor in

$$d(a^*c)^p a^{m_2+m_1}cv_l\bar{c}d(a^*c)^p a^{m_2}(a^*c)^{q+(h-1)r}(\bar{c} \mid \bar{d}).$$

Since this factor is in $\mathrm{fac}(H)$, Lemma 18, part (i), implies

$$\begin{aligned} m_2 + m_1 &= p + q + (h-1)r + 1 \\ &= (h-1)r + l + 1. \end{aligned} \tag{24}$$

Note that the string w_j^3 has a factor in

$$d(a^*c)^p a^{m_2+m_1}cv_l\bar{c}d(a^*c)^p a^{m_2+m_1}cv_l\bar{c}.$$

Since we assumed that $w_j^3 \in \mathrm{fac}(H)$, this factor is also in $\mathrm{fac}(H)$. By Lemma 19, $v_l\bar{c}$ has a prefix that belongs to $(a^*c)^l\bar{c}$. By Lemma 18, part (i), then, we have

$$\begin{aligned} m_2 + m_1 &= p + 1 + l + 1 \\ &= p + l + 2. \end{aligned} \tag{25}$$

From (24) and (25), we get

$$(h-1)r = p + 1.$$

Since $p \geq 0$, this implies $r \neq 0$ and

$$h = \frac{p+1}{r} + 1,$$

which shows that h is unique.

Case 2. $\psi(w_j) \in \bar{d}^*\bar{c}R$. This case is exactly symmetric to the preceding case.

Case 3. $\psi(w_j) \in V\bar{c}dc^+$. We can use Lemma 14 to infer

$$w_j \in v\bar{c}d(a^*c)^+a^*,$$
$$u_jw_{j+1}u_{j+1}\dots w_ku_k \in (a^*c)^*\bar{c}\hat{\Sigma}^*$$

for some string $v \in \text{fac}(\{\, v_n \mid n \in \mathbb{N}\,\})$ such that $\psi(v) \in V$. By Lemma 21, part (i), we can write

$$w_j \in v\bar{c}da^{l_1+l_2}c\dots a^{l_1+1}ca^{m_1},$$
$$u_jw_{j+1}u_{j+1}\dots w_ku_k \in a^{m_2}ca^{l_1-1}c\dots ca^1c\bar{c}\hat{\Sigma}^*$$
$$\subseteq (a^*c)^{l_1}\bar{c}\hat{\Sigma}^*.$$

for some $l_1, m_1, m_2 \geq 0$ and $l_2 \geq 1$ such that $m_1 + m_2 = l_1$. Similarly to Case 1, there must be some $r \geq 0$ such that

$$u_jw^i_{j+1}u_{j+1}\dots w^i_ku_k \in (a^*c)^{l_1+(i-1)r}\bar{c}\hat{\Sigma}^*$$

for all $i \geq 1$. Then $w^h_ju_jw^h_{j+1}u_{j+1}\dots w^h_ku_k$ has a factor in

$$(c \mid d)a^{l_1+1}ca^{m_1}v\bar{c}da^{l_1+l_2}c\dots a^{l_1+1}ca^{m_1}(a^*c)^{l_1+(h-1)r}\bar{c}\hat{\Sigma}^*$$
$$\subseteq (c \mid d)a^{l_1+1}ca^{m_1}v\bar{c}d(a^*c)^{l_2+l_1+(h-1)r}\bar{c}\hat{\Sigma}^*. \qquad (26)$$

This factor is in $\text{fac}(H)$. Note that the above inclusion holds even when $l_1 = r = 0$, since $l_1 = 0$ implies $m_1 = 0$.

We show that $a^{m_1}v \in H$. Recall $\psi(v) \in V$ and $v \in \text{fac}(\{\, v_n \mid n \in \mathbb{N}\,\})$. If $\psi(v) = \varepsilon$, then $v \in (a \mid b)^*$, but since $ca^{m_1}v\bar{c} \in \text{fac}(H)$, Lemma 14 implies $a^{m_1}v = \varepsilon \in H$. If $\psi(v) \in LR$, Lemma 22 implies that $a^{m_1}v \in a^*cv_l\bar{c}dv_l\bar{d}b^*$ for some l. Since $ca^{m_1}v\bar{c} \in \text{fac}(H)$, it follows from Lemma 19 and Lemma 18, part (i) and (ii), that $a^{m_1}v = a^{l+1}cv_l\bar{c}dv_l\bar{d}b^{l+1} = v_{l+1} \in H$.

So the set (26) is included in

$$(c \mid d)a^{l_1+1}cH\bar{c}d(a^*c)^{l_2+l_1+(h-1)r}\bar{c}\hat{\Sigma}^*.$$

Since there is an element of $\text{fac}(H)$ belonging to this set, we obtain by Lemma 18, part (i)

$$l_1 + 1 = l_2 + l_1 + (h-1)r + 1.$$

Since $h > 1$, $r \geq 0$ and $l_2 \geq 1$, this is a contradiction.

Case 4. $\psi(w_j) \in \bar{d}^+\bar{c}dV$. This case is exactly symmetric to the preceding case. □

Corollary 25 *The language H is not iterative.*

Corollary 26 *There is a 3-MCFL that is not k-iterative for any k.*

4 Conclusion

We have proved that the language H is a 3-MCFL that is not iterative. A simple consequence of this theorem is that if $\mathscr{C}$ is a subclass of the class MCFL of multiple context-free languages and $\mathscr{C}$ consists entirely of iterative sets, then the language H does not belong to $\mathscr{C}$ and hence $\mathscr{C}$ must be a proper subclass of MCFL.

Kanazawa and Salvati [8] showed that the class $\mathrm{MCFL}_{\mathrm{wn}}$ of well-nested multiple context-free languages is properly included in MCFL, and in particular, the language $\{ w\#w \mid w \in D_2^* \}$ belongs to $\mathrm{MCFL} - \mathrm{MCFL}_{\mathrm{wn}}$. Since every language in $\mathrm{MCFL}_{\mathrm{wn}}$ is k-iterative for some k, the language H serves as a further witness to the separation of MCFL and $\mathrm{MCFL}_{\mathrm{wn}}$.

Another subclass of MCFL that only contains languages that are k-iterative for some k is the class of languages in Weir's *control language hierarchy* [16,12,7]. As far as we know, it has been an open question whether the inclusion of the control language hierarchy in the class of multiple context-free languages is proper. The language H serves as a witness to the properness of the inclusion.

Corollary 27 *There is a 3-MCFL that does not belong to Weir's control language hierarchy.*

References

1. Berstel, J., Boasson, L.: Context-free languages. In: J. van Leeuwen (ed.) Handbook of Theoretical Computer Science, vol. B, pp. 59–102. Elsevier, Amsterdam (1990)
2. Greibach, S.A.: Hierarchy theorems for two-way finite state transducers. Acta Informatica **11**, 89–101 (1978)
3. Greibach, S.A.: One-way finite visit automata. Theoretical Computer Science **6**, 175–221 (1978)
4. Groenink, A.V.: Mild context-sensitivity and tuple-based generalizations of context-free grammar. Linguistics and Philosophy **20**(6), 607–636 (1997)
5. Groenink, A.V.: Surface without Structure. Ph.D. thesis, University of Utrecht (1997)
6. Kanazawa, M.: The pumping lemma for well-nested multiple context-free languages. In: V. Diekert, D. Nowotka (eds.) Developments in Language Theory: 13th International Conference, DLT 2009, *Lecture Notes in Computer Science*, vol. 5583, pp. 312–325. Springer, Berlin (2009)
7. Kanazawa, M., Salvati, S.: Generating control languages with abstract categorial grammars. In: Preliminary Proceedings of FG-2007: The 12th Conference on Formal Grammar (2007)
8. Kanazawa, M., Salvati, S.: The copying power of well-nested multiple context-free grammars. In: A.H. Dediu, H. Fernau, C. Martín-Vide (eds.) Language and Automata Theory and Applications, Fourth International Conference, LATA 2010, *Lecture Notes in Computer Science*, vol. 6031, pp. 344–355. Springer, Berlin (2010)
9. Kasami, T., Seki, H., Fujii, M.: Generalized context-free grammars, multiple context-free grammars and head grammars. Tech. rep., Osaka University (1987)
10. Kracht, M.: The Mathematics of Language. Mouton de Gruyter, Berlin (2003)
11. Michaelis, J.: On Formal Properties of Minimalist Grammars. Linguistics in Potsdam 13. Universitätsbibliothek, Publikationsstelle, Potsdam. Ph.D. thesis, ISBN 3-935024-28-2
12. Palis, M.A., Shende, S.M.: Pumping lemmas for the control language hierarchy. Mathematical Systems Theory **28**(3), 199–213 (1995)
13. Radzinski, D.: Chinese number-names, tree adjoining languages, and mild context-sensitivity. Computational Linguistics **17**(3), 277–299 (1991)
14. Seki, H., Matsumura, T., Fujii, M., Kasami, T.: On multiple context-free grammars. Theoretical Computer Science **88**(2), 191–229 (1991)
15. Vijay-Shanker, K., Weir, D.J., Joshi, A.K.: Characterizing structural descriptions produced by various grammatical formalisms. In: 25th Annual Meeting of the Association for Computational Linguistics, pp. 104–111 (1987)

16. Weir, D.J.: A geometric hierarchy beyond context-free languages. Theoretical Computer Science **104**(2), 235–261 (1992). DOI 10.1016/0304-3975(92)90124-X