Classification of strong codes

Yoshiyuki Kunimochi
Shizuoka Institute of Science and Technology

abstract Deletion and insertion are interesting and common operations which often appear in text editing. A language $L \subset A^*$ closed under the both operations forms a free submonoid of A^*. Its base C is identical to a strong code, which is a kind of bifix code. The class of hyper-strong codes is a subclass of strong codes. The star closure C^* of a hyper-strong code C is closed under scattered deletion and insertion. In the last part of this paper, we show that intercode is not a maximal code by using completeness of a bifix code.

1 Preliminaries

Let A be a finite nonempty set of *letters*, called an *alphabet* and let A^* be the free monoid generated by A under the operation of catenation with the identity called the *empty word*, denoted by 1. We call an element of A^* a word over A. The free semigroup $A^* \setminus \{1\}$ generated by A is denoted by A^+. The catenation of two words x and y is denoted by xy. The *length* $|w|$ of a word $w = a_1a_2 \ldots a_n$ with $a_i \in A$ is the number n of occurrences of letters in w. Clearly, $|1| = 0$. For a letter a in A, we let $|w|_a$ denote the number of occurrences of a in w. We denote $\{a \in A \,|\, xay \in L, x, y \in A^*\}$ by $\mathrm{alph}(L)$.

A word $u \in A^*$ is a *prefix*(resp. *suffix*) of a word $w \in A^*$ if there is a word $x \in A^*$ such that $w = ux$(resp. $w = xu$). A word $u \in A^*$ is a *factor* of a word $w \in A^*$ if there exist words $x, y \in A^*$ such that $w = xuy$. Then a prefix (a suffix or a factor) u of w is called *proper* if $w \neq u$.

A subset of A^* is called a *language* over A. A nonempty language C which is the set of free generators of some submonoid M of A^* is called a *code* over A. Then C is called the *base* of M and coincides with the minimal set $Min(M) = (M \setminus 1) \setminus (M \setminus 1)^2$ of generators of M. This is equivalent to the condition that $wC^* \cap C^* \neq \emptyset$ and $C^* \cap C^*w \neq \emptyset$ imply $w \in C^*$ for any $w \in A^*$. A nonempty language C is called a *prefix* (or *suffix*) code if $u, uv \in C$ (resp.$u, vu \in C$) implies $v = 1$. C is called a *bifix* code if C is both a prefix code and a suffix code. A nonempty language C is called a *hypercode* code if $u_1u_2 \ldots u_nu_{n+1}, u_1v_1u_2v_2 \ldots u_nv_nu_{n+1} \in C$ implies $v_1v_2 \ldots v_n = 1$. The language $A^n = \{w \in A^* \,|\, |w| = n\}$ with $n \geq 1$ is called a *full uniform* code over A. A code (resp. prefix code, suffix code, bifix code)C is called *maximal* (resp. **prefix-maximal**, **suffix-maximal**, **bifix-maximal**) if $C \cup \{w\}$ is not a code (resp. prefix code, suffix code, bifix code) for any $w \in A^* \setminus C$. A nonempty subset of A^n is called a *uniform* code over A. The symbols $\subset$ and $\subsetneq$ are used for a subset and a proper subset respectively.

A language L over A is called reflexive (resp. commutative) if $uv \in L$ implies $vu \in L$ (resp. $xuvy \in L$ implies $xvuy \in L$). The conjugacy class $cl(w)$ of a word w is the set $\{vu|w = uv\}$ and $w' \in cl(w)$ is called a conjugate of w.

Let N be a submonoid of a monoid M. N is right unitary (in M) if $u, uv \in N$ implies $v \in N$. Left unitary is defined in a symmetric way. The submonoid N of M is biunitary if it is both left and right unitary. Especially when $M = A^*$, a submonoid N of A^* is right unitary (resp. left unitary, biunitary) if and only if the minimal set $min(N) \stackrel{\text{def}}{=} (N \setminus 1) \setminus (N \setminus 1)^2$ of generators of N, namely the base of N, is a prefix code (resp. a suffix code, a bifix code) ([BP85] p.46).

Let L be a subset of a monoid M, the congruence $P_L = \{(u, v) \,|\, \text{for all } x, y \in M,\ xuy \in L \iff xvy \in L\}$ on M is called the *principal congruence*(or *syntactic congruence*) of L. We write $u \equiv v\ (P_L)$ instead of $(u, v) \in P_L$. The monoid M/P_L is called the *syntactic monoid* of L, denoted by Syn(L). The morphism ϕ_L of M onto Syn(L) is called the *syntactic morphism* of L. $\phi_L(w)$ is often denoted by $[w]_L$ or $[w]$. In particular when $M = A^*$, a language $L \subset A^*$ is regular if and only if Syn(L) is finite.

2 Strong Codes

A strong code C is the nonempty base of the identity $\bar{1}_L$ in the syntactic monoid $Syn(L)$ of some language L. Then we introduce the definition of strong codes and their properties.

2.1 Strong codes and Hyper-strong codes

At first, we give the definition of strong codes.

DEFINITION 2.1 A code $C \subset A^+$ with $C \neq \emptyset$ is called a *strong* code if

$$\begin{array}{lll} \text{(i)} & x, y_1y_2 \in C^* & \Rightarrow \quad y_1xy_2 \in C^* \\ \text{(ii)} & x, y_1xy_2 \in C^* & \Rightarrow \quad y_1y_2 \in C^* \end{array}$$

Note that if a code C satisfies the condition (ii), we can easily check that C^* is biunitary ($uv, u \in C^*$ implies $v \in C^*$ and $uv, v \in C^*$ implies $u \in C^*$). Then, C is a bifix code. Indeed, $uv, u \in C$ implies $v \in C^*$ and thus $v = 1$ because C is a code. Therefore C is a prefix code. Similarly C is a suffix code.

DEFINITION 2.2 [Cao92] Let $x_1, \ldots, x_n, y_1, \ldots, y_n, y_{n+1} \in A^*$. Then, $C \subseteq A^+$ is called an n-**strong** code if

$$\begin{array}{lll} \text{(i)} & x_1 \ldots x_n, y_1 \ldots y_ny_{n+1} \in C^* & \Rightarrow \quad y_1x_1 \ldots y_nx_ny_{n+1} \in C^* \\ \text{(ii)} & x_1 \ldots x_n, y_1x_1 \ldots y_nx_ny_{n+1} \in C^* & \Rightarrow \quad y_1 \ldots y_ny_{n+1} \in C^* \end{array}$$

Moreover, C is a **hyper-strong code** if C is an n-strong code for all $n > 0$.

A strong code C is described as the base of the identity P_L-class $\bar{1}_L = \{w \in A^* \,|\, w \equiv 1(P_L)\} \neq \{1\}$ of the syntactic monoids Syn(L) of some language L.

PROPOSITION 2.1 [H.J91] Let $L \subset A^*$. Then $C = (\bar{1}_L \setminus 1) \setminus (\bar{1}_L \setminus 1)^2$ is a strong code if it is not empty. Conversely, if $C \subset A^+$ is a strong code, then there exists a language $L \subset A^*$ such that $\bar{1}_L = C^*$.

A relation ρ on the free submonoid C^* of A^* is defined as follows:

$u\rho v$ if and only if there exist $m \in C^+$ $x_1, x_2 \in A^*$ such that $u = x_1x_2$ and $v = x_1mx_2$.

Let $\bar{\rho}$ the reflexive and transitive closure of ρ.

DEFINITION 2.3 [Zha87] Let C be a strong code over A. The root of C is the set:

$$R(C) = \{c \in C^+ | \forall c_1 \in C^+ (c_1\bar{\rho}c) \rightarrow c_1 = c\}.$$

PROPOSITION 2.2 [Zha87] Let C be a strong code over A. The followings are equivalent:
(1) C is a group code.
(2) $Syn(C^*)$ is a group (and then C^* is a P_{C^*}-class).
(3) C^* is reflexive;
(4) $R(C)$ is reflexive.
(5) C **is a maximal code.**
(6) C is a prefix-maximal (or suffix-maximal) code.
(7) $C \cap aA^* \neq \emptyset$ for $\forall a \in A$;
(8) $C \cap A^*a \neq \emptyset$ for $\forall a \in A$;

PROPOSITION 2.3 [H.J91] Let C be a strong code over A. Then C is finite if and only if C is a full uniform code over A, i.e. $C = A^n$ for some n.

This proposition means that a finite strong code is a maximal code.

PROPOSITION 2.4 [Lon96] Let C be a code over $A = alph(C)$. The conditions (1)∼(5) are equivalent:
(1) C is a maximal hyper-strong code over A;
(2) C^* is commutative;
(3) C^* is a P_{C^*}-class, and $Syn(C^*)$ is an Abelean group.
(4) $R(C)$ is hypercode.
(5) $R(C)$ is commutative.

EXAMPLE 2.1 (1) A^n(a full uniform code) is a finite maximal (hyper-)strong code.
$Syn(A^n)$ is the cyclic group $\langle x | x^n = 1\rangle$ of order n
with $\phi_{A^{n*}}(a) = x$ for $\forall a \in A$.
(2) $C_1 = \{a\} \cup ba^*b$ is a regular maximal (hyper-)strong code.
$Syn({C_1}^*)$ is also the cyclic group $\langle x | x^2 = 1\rangle$ of order 2
with $\phi_{{C_1}^*}(a) = 1, \phi_{{C_1}^*}(b) = x$.
(3) $C_2 = ba^*b$ is a regular strong code which is not maximal.
$Syn({C_2}^*) = \{[1], [a], [b], [ab], [ba]\}$ is a monoid but not a group
with $\phi_{{C_2}^*}(a) = [a], \phi_{{C_2}^*}(b) = [b]$.
Note that $[a]^2 = [a], [b]^2 = [1], [ab]^2 = [ba]^2 = [a], [ba][b] = [b][ab] = [1]$.

EXAMPLE 2.2 **(semi-Dyck languages)** Let $A = \{(_i \,|\, 1 \le i \le r\}$ and $\bar{A} = \{)_i \,|\, 1 \le i \le r\}$ be disjoint alphabets. The semi-Dyck languge ${D'_r}^*$ on $A \cup \bar{A}$ is generated by a context-free grammar with the rules:

$$S \to (_iS)_iS, S \to 1.$$

For example $(_1)_1$ and $(_1(_1)_1)_1(_1)_1$ are in ${D'_1}^*$.

(1) The base $D'_r = min({D'_r}^*)$ is a strong code, **not maximal**, not a hyper-strong, especially D'_1 is only a hyper-strong code.
(2) The root $R(D'_r) = \{(_i)_i \,|\, 1 \le i \le r\}$.
(3) ${D'_r}^*$ is a simple language.
(4) $Syn({D'_1}^*)$ is the **bicycle monoid** generated by the tranformation x and y on the set of natural numbers below(Figure.2.1).

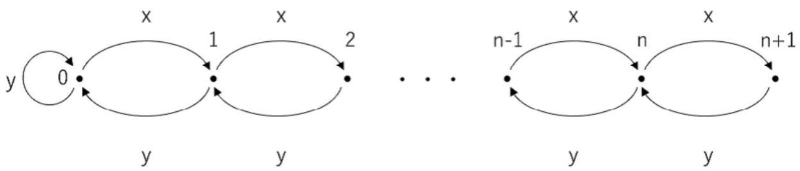

Figure 1. bicyclemonoid

C	strong code	hyper-strong code
Finite	A^n for some $n > 0$	
Regular	$C_2 = ba^*b$ †	$C_1 = \{a\} \cup ba^*b$
		$\{w \| \ \|w\|_a \equiv 2\|w\|_b \pmod 3\}$
Context-free	Semi-Dyck code D'_r †	Dyck code D_r

† = strong codes which are not maximal.

Figure 2. Classification of strong codes

EXAMPLE 2.3 (Dyck languages) Let $A = \{(_i \,|\, 1 \leq i \leq r\}$ and $\bar{A} = \{)_i \,|\, 1 \leq i \leq r\}$ be disjoint alphabets. The Dyck language $D_r{}^*$ on $A \cup \bar{A}$ is generated by a context-free grammar with the rules:

$$S \rightarrow (_iS)_iS, S \rightarrow)_iS(_iS, S \rightarrow 1.$$

(1) The base $D_r = min(D_r{}^*)$ is a **maxial** strong code, especially D_1 is only a hyper-strong code.
(2) The root $R(D_r) = \{(_i)_i,)_i(_i | 1 \leq i \leq r\}$.
(3) $D_r{}^*$ is a simple language.
(4) Syn($D_r{}^*$) is a free group $F(A)$. Especially Syn($D_1{}^*$) is only an Abelean group $(\mathbf{Z}, +)$.

The following corollary and proposition give a necessary condition and a sufficient condition that a strong code has a finite root, respectively.

COROLLARY 2.1 [Zha87] Let C be a strong code over A. If the root $R(C)$ is finite, then C^* is context-free.

PROPOSITION 2.5 [Zha87] Let C be a strong code over A. If C is regular, then the root $R(C)$ is finite.

Zhang conjectured that a strong code has a finite root if and only if it is a simple language. Whereas Harging-Smith[HS83] proved the following theorem in 1973. In the theorem, Let $\pi = \langle X; R \rangle$ be a finitely generated presentation of a group G, and $A = X \cup X^{-1}$ be the set of generators and their inverses. The word problem $\mathrm{WP}(\pi)$ of π is the set of all words on A which are equal to the identity. The reduced word problem $\mathrm{WP}_0(\pi)$ of π is the set $\mathrm{WP}(\pi) \backslash \mathrm{WP}(\pi)A^+$. The set $\mathrm{W}(\pi)$ of *irreducible words* is the set $\mathrm{WP}(\pi) \setminus A^+\mathrm{WP}(\pi)A^+$

DEFINITION 2.4 A context-free grammar $G = (V, \Sigma, P, S)$ in Greibach normal form is said to be a *simple grammar* if for all $A \in V - \Sigma, a \in \Sigma$, and $\alpha, \beta \in V^*$,

$$A \to a\alpha, \text{ and } A \to a\beta \text{ imlpy } \alpha = \beta.$$

A simple language is a language generated by a simple grammar.

THEOREM 2.1 [HS83] The reduced word problem $\mathrm{WP}_0(\pi)$ of a finitely generated group presentation π is a simple language if and only if the set $W(\pi)$ of irreducible words is finite.

2.2 Insertion and Deletion

Let L be a language over A. A language L is called **ins-closed** if $u = u_1u_2 \in L$ and $v \in L$ imply $u_1vu_2 \in L$. A language L is called **del-closed** if $u = u_1vu_2 \in L$ and $v \in L$ imply $u_1u_2 \in L$ [IKT97].

Let L be a del-closed language. Then, Since L is biunitary, the minimal set $C = min(L)$ of generators of L is a bifix code and $L = C^*$.

Let L be an ins-closed language. Then, $1 \in L$ and $L^2 \subset L$ implies Since L is a submonoid of A^*.

PROPOSITION 2.6 Let $L \neq \emptyset$ be an ins-closed and del-closed language over A. Then $L = C^*$ for some strong code C.

Proof) As we stated above, L is a submonoid of A^* and its minimal set C of generators is a (bifix) code. C satisfies the conditions of a strong code.

3 Intercodes

In the last section, we show that any intercode of index m is not maximal. At first, we give the definition of an intercodes of index m.

DEFINITION 3.1 [Yu05] A language C over A is called an **intercode of index** $m\,(\geq 1)$ if $C^{m+1} \cap A^+C^mA^+ = \emptyset$. ∎

EXAMPLE 3.1 (1) $I_1 = \mathrm{b}a^+\mathrm{b}$ is an intercode of index 1(infix and not maximal).
(2) $I_2 = \{ba, \mathrm{c}ba\mathrm{d}\}$ is an intercode of index 2 but is not an intercode of index 1(bifix, not maximal).
(3) $I_3 = \{ba, ba^2, \cdots, ba^{m-1}, \mathrm{c}baba^2 \cdots ba^{m-1}\mathrm{d}\}$ is an intercode of index m but is not an intercode of index k with $k < m$ (bifix, not maximal).

PROPOSITION 3.1 (1) An intercode C of index $m\,(\geq 1)$ is a bifix code and thin.
(2) A full uniform code is **not** an intercode of index m for any $m \geq 1$.

$\because$) (1) If $u, ux \in C \subseteq A^+, x \in A^*$, then $u^m \in C^m, uu^mx \in C^{m+1}$, and thus $x = 1$. C is a prefix code. Similarily C is a suffix code and thus a bifix code. For any $u \in C$, $C \cap A^*u^{m+2}A^* = \emptyset$. Therefore, C is thin. (2) Since $(A^n)^{m+1} \cap A^+(A^n)^mA^+ \neq \emptyset$, A^n is not a full uniform code.

∎

PROPOSITION 3.2 [BPR10] If X and Y are prefix codes(resp. prefix-maximal codes, suffix-maximal codes), then XY is a prefix code(resp. prefix-maximal code, suffix-maximal).

PROPOSITION 3.3 [BPR10] Let X be a **thin** subset of A^+, The following conditions are equivalent:
(i) X is a maximal code and bifix.
(ii) X is a bifix-maximal code.
(iii) X is a prefix-maximal code and a suffix-maxial code.
...

COROLLARY 3.1 If C is a bifix-maximal code, then C^m is a bifix-maximal code.

PROPOSITION 3.4 [Lal79], p.235 If C is a complete bifix code over in A^*, and if $C \neq A^n$ for all $n \geq 1$, then there exists $c \in C$ such that $c \in A^+CA^+$.

PROPOSITION 3.5 For any positive integer m, no intercode of index m is a maximal code.

$\because$) Let C be an intercode of index m. Suppose that C is a maximal code. Then by Corollary 3.1, C^m is a **bifix-maximal** code. Since C is an intercode of index m, $C^m \cap A^+C^mA^+ = \emptyset$. By **Proposition 3.4**, C^m and thus C must be full uniform codes but this is a contradiction to **Proposition3.1**. Therefore C is not maximal.

∎

References

[BP85] Jean Berstel and Dominique Perrin, *Theory of codes*, Academic Press, 1985.

[BPR10] Jean Berstel, Dominique Perrin, and Christophe Reutenauer, *Codes and automata*, no. 129, Cambridge University Press, 2010.

[Cao92] WL Cao, *Hyper-strong codes, their properties structures, j*, Lanshou Uni. 28 (1992) (1992), 10–15.

[Har78] Michael A Harrison, *Introduction to formal language theory*, Addison-Wesley Longman Publishing Co., Inc., 1978.

[H.J91] H.J.Shyr, *Free monoids and languages*, Lecture Notes, Hon Min book Company, Taichung, Taiwan, 1991.

[HS83] Robert H Haring-Smith, *Groups and simple languages*, Transactions of the American Mathematical Society **279** (1983), no. 1, 337–356.

[IJST91] Masami Ito, Helmut Jürgensen, Huei-Jan Shyr, and Gabriel Thierrin, *Outfix and infix codes and related classes of languages*, Journal of Computer and System Sciences **43** (1991), no. 3, 484–508.

[IKT97] Masami Ito, Lila Kari, and Gabriel Thierrin, *Insertion and deletion closure of languages*, Theoretical computer science **183** (1997), no. 1, 3–19.

[Kas75] Takumi Kasai, *A universal context-free grammar*, Information and Control **28** (1975), no. 1, 30–34.

[Kun16] Yoshiyuki Kunimochi, *Some properties of extractable codes and insertable codes*, International Journal of Foundations of Computer Science **27** (2016), no. 03, 327–342.

[Lal79] Gérard Lallement, *Semigroups and combinatorial applications*, John Wiley & Sons, Inc., 1979.

[LJD97] Dong Yang Long, Ma Jian, and Zhou Duanning, *Structure of 3-infix-outfix maximal codes*, Theoretical Computer Science **188** (1997), no. 1-2, 231–240.

[Lon92] Dongyang Long, *On the structure of some group codes*, Semigroup Forum, vol. 45, Springer, 1992, pp. 38–44.

[Lon96] ______, *On group codes*, Theoretical computer science **163** (1996), no. 1-2, 259–267.

[NB21] Carl-Fredrik Nyberg-Brodda, *The word problem for one-relation monoids: a survey*, Semigroup Forum, vol. 103, Springer, 2021, pp. 297–355.

[RS97] Grzegorz Rozenberg and Arto Salomaa, *Handbook of formal languages: Volume 1 word, language, grammar*, Springer Science & Business Media, 1997.

[RT79] CM Reis and Gabriel Thierrin, *Reflective star languages and codes*, Information and Control **42** (1979), no. 1, 1–9.

[Yu05] SS Yu, *Languages and codes. lecture notes, department of computer science*, 2005.

[Zha87] Louxin Zhang, *Rational strong codes and structure of rational group languages*, Semigroup forum, vol. 35, Springer, 1987, pp. 181–193.

[ZQ93] Liang Zhang and Weide Qiu, *Decompositions of recognizable strong maximal codes*, Theoretical computer science **108** (1993), no. 1, 173–183.

Yoshiyuki Kunimochi
Shizuoka Institute of Science and Technology
Toyosawa 2200-2, Fukuroi-shi, Shizuoka 437-8555,
JAPAN
Email: kunimochi.yoshiyuki@sist.ac.jp