On Regularity and Roots 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 called a strong code, that is, $L = C^*$. The language L is regular if and only if its base C is regular. Then, we prove in another way that the syntactic monoid of L becomes a finite group. This gives us many examples of regular strong codes. We also investigate the relation between strong codes and groups.

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.

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. 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. The language $A^n = \{w \in A^* \,|\, |w| = n\}$ with $n \geq 1$ is called a *full uniform* code over A. 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.

We denote $\{a \in A \,|\, xay \in L, x, y \in A^*\}$ by $\mathrm{alph}(L)$. A language L over A is called reflexive if $uv \in L$ implies $vu \in L$. 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 $N_0 = (N \setminus 1) \setminus (N \setminus 1)^2$ of generators of N, namely the base of N, is a prefix code (resp. a suffix code, a bifix code) ([1] 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 $\mathrm{Syn}(L)$. The morphism σ_L of M onto $\mathrm{Syn}(L)$ is called the *syntactic morphism* of L. $\sigma_L(w)$ is denoted by $\bar{w}_L$. In particular when $M = A^*$, a language $L \subset A^*$ is regular if and only if Syn(L) is finite([1] p.46).

2 Strong Codes

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

2.1 definitions

At first, we give the definition of strong codes.

DEFINITION 2.1 [4] A code $C \subset A^+ \setminus \{\emptyset\}$ is called a *strong* code if

$$\text{(i)}\; x, y_1y_2 \in C^* \implies y_1xy_2 \in C^*$$
$$\text{(ii)}\; x, y_1xy_2 \in C^* \implies y_1y_2 \in C^*$$

Here extractable codes and insertable codes are introduced below.

DEFINITION 2.2 Let $C \subset A^+\setminus\{\emptyset\}$ be a code. Then, C is called an insertable (or extractable) code if C satisfies the condition (i)(or (ii)).

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

PROPOSITION 2.1 [4] 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 $\bar{1}_L = C^*$.

Moreover if a strong code C is finite, the following proposition holds.

PROPOSITION 2.2 [4] Let C be a finite strong code over A and $B = \text{alph}(C)$. Then, $C = B^n$ for some positive integer n, that is, C is a full uniform code over B.

EXAMPLE 2.1 (1) A singleton $\{w\}$ with $w \in \{a\}^+$ is a strong code. $\{w\}$ with $w \in A^+ \setminus \bigcup_{a\in A}\{a\}^+$ is not a strong code but it is an extractable code. Therefore there exist finite extractable codes which are not full uniform codes.
(2) The conjugacy class $cl(ab)$ of ab is an extractable code but not a strong code.
(3) $\{a^nb^n \mid n \text{ is an integer}\}$ is an (context-free) extractable code but not a strong code.
(4) a^*b and ba^* are (regular) insertable codes but not strong codes.

Note that when C satisfies the condition (ii), we can easily check that C^* is biunitary(and thus free). Indeed, $uv = 1uv, u \in C^*$ implies $v = 1v \in C^*$ and $uv = uv1, v \in C^*$ implies $u = 1u \in C^*$. Then the minimal set $C = (C^* \setminus 1) \setminus (C^* \setminus 1)^2$ of generators of C^* becomes a bifix code. Therefore both strong codes and extractable codes are necessarily bifix codes.

Remark that an insertable submonoid M of A^*, the minimal set of generators of M is not necessarily a code. For example, If $C = \{a^2, a^3\}$, then the submonoid C^* is insertable but its minimal set C of generators is not necessarily a code.

PROPOSITION 2.3 [18] Let C be a code over A. Then the following conditions are equivalent:
(1) C^* is reflexive;
(2) C is a maximal strong code over A;
(3) C^* is a P_{C^*}-class, $Syn(C^*)$ is a group.

Note that the condition (2) is equivalent to the following condition (2'):

(2') C is a strong code over A and $alph(C) = A$.

Indeed, if $a \in A \backslash alph(C)$, then $C \cup \{a\}$ is a code. This contradicts to the condition (2). Hence $alph(C) = A$. Conversely, suppose the condition (2'), that is $A = alph(C)$. We show that $C \cup \{w\}$ with any $w = a_1a_2 \ldots a_k \notin C (a_i \in A, 1 \leq i \leq k)$ cannot be a code. For any $a_i \in A$, $a_iy_i \in C$ for some $y_i \in A^*$ because C is reflexive. Therefore $w(y_k \ldots y_2y_1) = a_1a_2 \ldots a_ky_k \ldots y_2y_1 = c_1c_2 \ldots c_m \in C^*$ for some $c_j \in C (1 \leq j \leq m)$. Since C^* is reflexive again, $(y_k \ldots y_2y_1)w = c'_1c'_2 \ldots c'_n \in C^*$ for some $c'_j \in C (1 \leq j \leq n)$. Therefore $c_1c_2 \ldots c_mw = wc'_1c'_2 \ldots c'_n \in C^*$. This proves that $C \cup \{w\}$ is not a code.

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 $u_1u_2 \in L$ [6].

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 2.4 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. ∎

2.3 Roots of Strong Codes

Let L be a strong code over A. We define a relation ρ on the free submonoid C^* of A^* 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 [18] 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 2.5 [18] Let C be a strong code over A. Then the following conditions are equivalent:

(1) C is a maximal strong code;

(2) $R(C)$ is reflexive;

(3) $R(C) = \{w \in C|$ every conjugate w' of w is in $C\}$.

PROPOSITION 2.6 [18] Let C be a strong code over A. If the root $R(C)$ is finite, the there exist a Dyck language $D_k \subset (A_1)^*$ and a homomorphism $f : (A_1)^* \to A^*$ such that $C^* = f(D_k)$

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

PROPOSITION 2.7 [18] 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[3] proved the following theorem in 1973. In the theorem, Let $\pi =< A; R >$ be a finitely generated presentation of a group G, and $\Sigma = A \cup A^{-1}$ be the set of generators and their inverses. The word problem $WP(\pi)$ of π is the set of all words on Σ which are equal to the identity. The reduced word problem $WP_0(\pi)$ of π is the set $WP(\pi) \setminus WP(\pi)\Sigma^+$. The set $W(\pi)$ of irreducible words is the set $WP(\pi) \setminus \Sigma^+ WP(\pi)\Sigma^+$

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 $A \in N, a \in \Sigma$, 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 [3] The reduced word problem $WP_0(\pi)$ of a finitely generated group presentation π is a simple language if and only if the set of irreducible words $W(\pi)$ is finite.

To prove the conjecture, It remains to check that for any finitely generated presentation $\pi =< A; R >$ of a group G with $WP(\pi) \neq \emptyset$,

· The correspondence between strong codes and reduced word problems.

· $WP_0(\pi)$ is a strong codes and $W(\pi)$ is its root.

· $WP_0(\pi) \cap A^*$ is a strong codes and $W(\pi) \cap A^*$ is its root.

EXAMPLE 2.2 Let Σ be an alphabet and let $\bar{\Sigma}$ be its copy. The Dyck language ${D_\Sigma}^*$ over Σ is generated by the context-free grammar $(\{S, T\}, \Sigma \cup \bar{\Sigma}, P, S)$, where

$$S \to \varepsilon, S \to TS, T \to aS\bar{a}\,(a \in \Sigma).$$

${D_\Sigma}^*$ is a free submonoid of $(\Sigma \cup \bar{\Sigma})^*$ and its base D_Σ is a strong code over $\Sigma \cup \bar{\Sigma}$. If $|\Sigma| = n$, then D_Σ is often denoted by D_n.

D_n is not a regular language. The root of D_n is the set $R(D_n) = \{a\bar{a} \mid a \in \Sigma\}$

EXAMPLE 2.3 The language $L = \{w \mid |w|_a = |w|_b\}$ over $A = \{a, b\}$ is ins-closed and del-closed. L is a free submonoid of A^*. Its base $C = min(L)$ is a maximal strong code of even length over A. The root $R(C)$ of C is the set $R(C) = \{ab, ba\}$

3 regular strong codes

We show that regular strong code is a maximal bifix code by another approach.

THEOREM 3.1 Let L be a regular ins-closed and del-closed language and $C = min(L)$ be the minimal set of generators of L. N be the number of states in a minimal automaton recognizing L. Then the following statements hold.
(1) For any $x \in alph(L)^*$, $x^n \in L$ for some positive integer $n \leqq N$.
(2) Let $m \in M = Syn(L)$, $m^n = 1$ for some n that is M is a finite group.

LEMMA 3.1 Let $L, C = min(L)$ and N are the same as those in the theorem. $uv \in L$ implies $u^m \in L$ for some $0 < m \leqq N$

Proof) Let $A = (Q, \Sigma, \delta, s_0, F)$ be a minimal automaton recognizing L. $\delta(s_0, u^s) = \delta(s_0, u^t)$ for some s, t $(0 \leqq s < t \leqq N)$ since $|Q| = N$. $u^s v^s \in L$ because L is ins-closed and del-closed. Setting $0 < i = t - s \leqq N$, $u^{s+i}v^s = u^i(u^s v^s) \in L$. Again since L is ins-closed and del-closed, $u^i \in L$. ∎

Proof of theorem 3.1) (1) Let $x \in alph(L)^*$ be an arbitrary word. Let $a \in alph(L)$, that is $uav \in L$. By Lemma 1, $u^n \in L$ for some n. Since L is ins-closed and del-closed, $u^n(av)^n \in L$. $a(vav \cdots av) \in L$ holds. We get $a^i \in L (0 < i \leqq N)$ again by Lemma 1.

$$a_1 a_2 \cdots a_r (a_r)^{i_r - 1} a_{r-1}^{i_{r-1}-1} \cdots {a_1}^{i_1 - 1} \in L.$$

By Lemma 1, $x^n \in L$ for $0 < n \leqq N$.

(2) Let $M = Syn(L)$ the syntactic monoid of L and $\phi : A^* \to Syn(L), u \mapsto \bar{u}$ the syntactic morphism. Since L is regular, M is finite. For any $m \in Syn(L)$, there exists $x \in alph(L)^*$ such that $\phi(x) = \bar{x} = m$. By (1), $x^n \in L$. $\bar{x}^n = \bar{1}$. Therefore $\bar{x}$ has an inverse element $\bar{x}^{n-1}$. Hence M is a finite group. ∎

COROLLARY 3.1 Suppose that $L, C = min(L)$ and N are the same as those in the theorem. Then, C is a strong code.

Proof) We show C is a maximal prefix code. C is a bifix code because L is biunitary. Let $x \in alph(L)^*$, $xx^{n-1} \in L = C^*$ for some n. This means maximality ∎

References

[1] J. Berstel and D. Perrin. *Theory of Codes*. Pure and Applied Mathematics. Academic Press, 1985.
[2] A. de Luca and S. Varricchio. *Finiteness and Regularity in Semigroups and Formal Languages*. Monographs on Theoretical Computer Science・An EATCS Series. Springer, July 1999.
[3] G. H. Haring-Smith. *Groups and simple languages*, volume 239. 9 1983.
[4] H.J.Shyr. Strong codes. *Soochow J. of Math. and Nat. Sciences*, 3:9–16, 1977.

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

[6] M. Ito, L. Kari, and G. Thierrin. Insertion and deletion closure of languages. *Theoretical Computer Science*, 183:3–19, 1997.

[7] J.M.Howie. *Fundamentals of Semigroup Theory*. London Mathematical Society Monographs New Series 12. Oxford University Press, 1995.

[8] Y. Kunimochi. Some properties of extractable codes and insertable codes. *International Journal of Foundations of Computer Science*, 27(3):327–342, 2016.

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

[10] D. Long. On the structure of some group codes. 45:38–44, 1992.

[11] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1983.

[12] T. Moriya and I. Kataoka. Syntactic congruences of codes. *IEICE TRANSACTIONS on Information and Systems*, E84-D(3):415–418, 2001.

[13] M.Petrich and G.Thierrin. The syntactic monoid of an infix code. *Proceedings of the American Mathematical Society*, 109(4):865–873, 1990.

[14] G. Rozenberg and A. Salomaa. *Handbook of Formal Languages, Vol.1 WORD, LANGUAGE, GRAMMAR*. Springer, 1997.

[15] G. Tanaka, Y. Kunimochi, and M. Katsura. Remarks on extractable submonoids. *Technical Report kokyuroku, RIMS, Kyoto University*, 1655:106–110, 6 2009.

[16] S. Yu. A characterization of intercodes. *International Journal of Computer Mathematics*, 36(1-2):39–45, 1990.

[17] S.-S. Yu. *Languages and Codes*. Tsang Hai Book Publishing Company, Taiwan, 2005.

[18] L. Zhang. Rational strong codes and structure of rational group languages. 35(1):181–193, 1987.

[19] L. Zhang and W. Qiu. Decompositions of recognizable strong maximal codes. 108:173–183, 1993.

[20] L. Zhang and W. Qiu. On group codes. 163:259–267, 1996.

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