Une formulation alternative de la conjecture de factorisation des codes

Jean-Marie Boë

Résumé

Le but de ce document est de donner une formulation équivalente de la conjecture de factorisation des codes finis en termes d'automates. Après avoir rappelé le problème et les définitions des codes et automates, nous montrons que la factorisation se ramène à un prolongement du code en un code baïonnette, puis le résultat principal. Un exposé très complet de la théorie des codes se trouve dans [BP85].

1 Introduction, définitions, rappels

La conjecture de factorisation des codes est un important problème ouvert de la théorie des codes. Nous établissons ici un lien entre cette conjecture, les codes "baïonnette" et les automates en pétales. A. Restivo [Rest77] a montré que les codes baïonnette sont factorisants. Nous utilisons ce résultat pour donner une formulation alternative de la conjecture de factorisation des codes.

Dans ce document A désigne un alphabet et A^* le monoïde libre qu'il engendre. Le produit de deux parties P et Q de A^* est *non-ambigu* si $m \in PQ$ entraîne l'existence d'un unique couple $(p, q) \in P \times Q$ tel que $m = pq$.

Définition 1.1 Un *code* est une partie X de A^* engendrant librement le sous-monoïde X^* : les produits X^n sont non-ambigus et $X^n \cap X^m = \emptyset$ pour $m \neq n$.

Received by the editors May 95.
Communicated by M. Boffa.
1991 *Mathematics Subject Classification :* 68Q45.
Key words and phrases : Codes à longueur variable, automates finis, séries formelles.

Exemple 1.2
- A^n est un code pour tout n.
- $\{a, ab, ba\}$ n'est pas un code car aba possède deux factorisations.
- $\{aaaa, ab, abaa, baa, b\}$ et $\{aa, ab, aab, abb, bb\}$ sont des codes.

La conjecture de la factorisation s'énonce alors :

Conjecture 1.3 *Tout code X fini et maximal pour l'inclusion est factorisant, i.e., il existe deux parties finies Q et P de A^* telles que :*

$$A^* = QX^*P \tag{1}$$

où les produits sont non-ambigus.

Exemple 1.4
- $A^* = (\bigcup_{0 \leq p < n} A^p)(A^n)^*\{\epsilon\}$,
- $A^* = \{\epsilon, b\}\{aa, ab, aab, abb, bb\}^*\{a, \epsilon\}$.

Les résultats les plus importants concernant cette conjecture proviennent de [Reu85] et [Sch65].

Pour $P \subseteq A^*$ on note $\underline{P}$ *la série (ou le polynôme) caractéristique* de P, élément de $\mathbb{Z}\langle\langle A \rangle\rangle$.

Lemme 1.5 *Soient P, Q, X des parties de A^**
- *Le produit PQ est non-ambigu si et seulement si $\underline{P}\,\underline{Q} = \underline{PQ}$.*
- *X engendre librement X^* si et seulement si $\underline{X^*} = 1/(1 - \underline{X})$.*

Preuve. Le premier point est immédiat, le second se déduit des égalités :

$$X^* = \bigcup_{n>=0} X^n \qquad 1/(1 - \underline{X}) = \sum_{n \geq 0} \underline{X}^n$$

■

Rappelons le résultat de Schützenberger concernant la maximalité des codes :

Proposition 1.6 *Soit X un code fini sur A :*

$$X \text{ est maximal pour l'inclusion} \quad ssi \quad \forall m \in A^* \quad A^*mA^* \cap X^* \neq \emptyset.$$

Nous supposerons désormais que les codes que nous considérons sont *finis et maximaux pour l'inclusion.*

On déduit du lemme 1.5 qu'un code est factorisant si et seulement si

$$\underline{A^*} = \underline{Q}\ \underline{X^*}\ \underline{P}$$

ce qui donne une écriture équivalente de (1) :

$$\underline{X} - 1 = \underline{P}(\underline{A} - 1)\underline{Q}. \tag{2}$$

Définition 1.7 Un *ensemble factorisant* est une partie X de A^* vérifiant :
- X est fini.
- il existe deux parties finies P et Q telles que $\underline{X} - 1 = \underline{P}(\underline{A} - 1)\underline{Q}$

Voici un lemme technique utilisé dans la preuve du résultat suivant et celle du résultat principal :

Lemme 1.8 *Soient X un ensemble factorisant et $q_0 \in Q$ (resp. $p_0 \in P$) un mot de longueur maximale de Q (resp. P) ; alors pour tout $p \in P$ (resp. $q \in Q$) il existe un mot $m \in A^*$ tel que $pmq_0 \in X$ (resp. $p_0mq \in X$).*

Preuve. Soit $p \in P$; si pour une lettre a, $paq_0 \in X$, le résultat est acquis ; sinon c'est que $paq_0 = p'q'$ pour $p' \in P, q' \in Q$; mais p' est de longueur supérieure à p car q_0 est de longueur maximale ; p est un début de p', on prouve donc par induction croissante le résultat. ∎

Proposition 1.9 *Un ensemble factorisant est un code.*

Preuve. Soit X un tel ensemble. Il vérifie donc l'égalité (2), dont on déduit :

$$\underline{A^*} = \underline{Q}\ (\sum_{n\geq 0} \underline{X^n})\ \underline{P}$$

d'où il vient que les produits X^n sont non-ambigus et les X^n sont disjoints deux à deux, ce qui est la condition de codicité. La maximalité de X se déduit du lemme 1.8 et de la proposition 1.7 : tout mot $m \in A^*$ se factorise en qxp avec $x \in X^*, q \in Q, p \in P$; mais il existe dans X des mots uq et pv, donc $umv \in X^*$. ∎

2 Automates

2.1 Automates non-ambigus

Définition 2.1 *Un automate non-ambigu $\mathcal{A}$* d'états E est un automate non-déterministe vérifiant de plus :
- non-ambiguïté : étant donnés deux états i et j, tout mot $m \in A^*$ est étiquette d'au plus un chemin de i à j.
- il possède un état e à la fois initial et final.

Proposition 2.2 *Un automate non-ambigu $\mathcal{A}$ reconnaît un sous-monoïde libre X^* ; X est l'ensemble des mots étiquettes d'un chemin menant de e à e sans passer par e dans l'automate $\mathcal{A}$.*

Preuve. Un mot m menant de e à e détermine une unique factorisation en mots de X conformément à la condition de non-ambigüité ; X^* est donc librement engendré par X. ∎

Pour s'assurer que la base X de ce sous-monoïde est un code (fini et maximal), il faut ajouter les conditions :
- l'automate $\mathcal{A}$ est émondé (fortement connexe) : pour tout couple d'états (i, j), il existe un chemin de i à j.
- l'automate $\mathcal{A}$ est complet : tout mot est étiquette d'au moins un chemin de l'automate.

– l'automate $\mathcal{A}$ ne possède pas de cycle évitant e.

Un tel automate sera appelée un *automate de code*.

Proposition 2.3 *Soit $\mathcal{A}$ un automate de code ; l'ensemble X des mots étiquettes d'un chemin menant de e à e sans passer par e est un code.*

Preuve. D'après la proposition 2.2, il reste à prouver que X est fini et maximal.

La finitude découle de l'absence de cycle évitant e : on ne peut avoir de chemin infini allant de e à e sans passer par e.

La maximalité découle de la proposition 1.6 : Soit $m \in A^*$; m est étiquette d'un chemin de i à j ($\mathcal{A}$ est complet) ; puisqu'il est émondé, il existe deux mots u et v tels que umv soit étiquette d'un chemin de e à e et donc $umv \in X^*$. ∎

Pour associer à tout code un automate de code, nous considérerons les automates en pétales.

2.2 Automates en pétales

Soit $X \subset A^*$ un code. *L'automate en pétales* $\mathcal{AP}$ de X a pour ensemble d'états :

$$E = \{u|v \mid u, v \in X, u \neq \epsilon \neq v\} \cup \{\epsilon|\epsilon\}$$

avec $e = \epsilon|\epsilon$. L'ensemble E est fini car X l'est. Les arcs d'étiquette a sont alors :

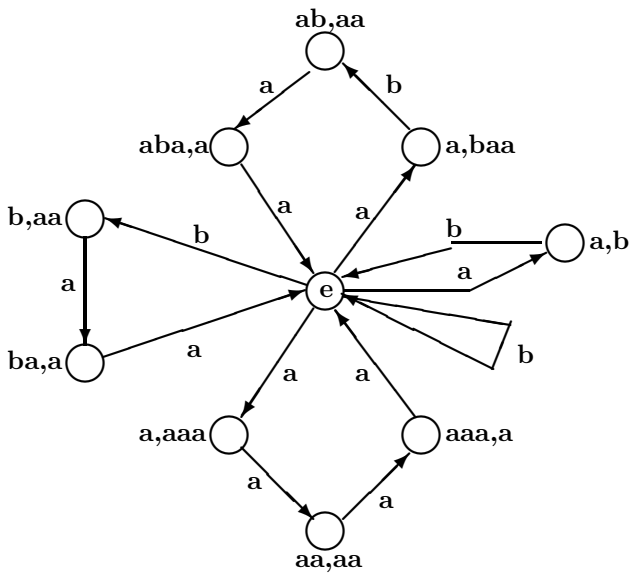
$$\{(ua|v, u|av) \mid uav \in X\} \cup \{(u|a, e) \mid ua \in X\} \cup \{(e, a|u) \mid au \in X\}$$

ce qui entraîne que $m \in A^*$ mène de $u|v$ à $r|s$ si $um \in X^*r$ et $ms \in vX^*$.

Remarquons que de tout état différent de e sort au plus un arc étiqueté d'une lettre donnée a et de même pour les arcs entrants.

– $\mathcal{AP}$ est non-ambigu : si un mot m est étiquette d'un chemin de i à j sans passer par e, il est unique d'après la remarque précédente ; si ce chemin passe par e, on le découpe en trois chemins menant respectivement de i à e, de e à e, de e à j ; ces trois chemins sont uniques : les deux extrêmes d'après l'affirmation précédente, l'intermédiaire car X est un code.
– $\mathcal{AP}$ est complet : X étant maximal, le théorème de Schützenberger affirme l'existence de deux mots u et v tels que $umv \in X^*$, c'est à dire que umv mène de e à e. Le mot m est donc étiquette d'un chemin.
– $\mathcal{AP}$ est un automate émondé : le mot vr mène de $u|v$ à $r|s$;
– $\mathcal{AP}$ ne possède pas de cycle évitant e : par construction même.

Exemple 2.4 Voici l'automate en pétales du code $\{aaaa, ab, abaa, baa, b\}$.



3 Prolongement, baïonnette et factorisation

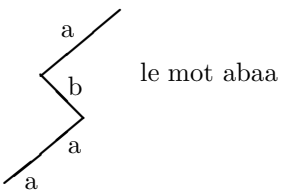
3.1 Baïonnette et factorisation

Définition 3.1 Un code X sur A est un code *baïonnette* s'il existe une lettre $b \in A$ pour laquelle on a :

$$X \subset (A\backslash b)^* b (A\backslash b)^* \cup (A\backslash b)^*$$

En d'autres termes, tout mot de X possède au plus une occurrence de la lettre b.

Exemple 3.2 L'ensemble $\{aaaa, ab, abaa, baa, b\}$ est un code baïonnette ; l'appellation "baïonnette" provient de la forme des chemins associés aux mots du code dans l'arbre des mots



Proposition 3.3 *Un code baïonnette est factorisant.*

Preuve. Soit b une lettre baïonnette d'un code baïonnette X. Comme X est maximal, la proposition 1.6 affirme que pour tout $m \in A^*$, $bmbb$ est complétable dans X^*, i.e. :

$$\exists u, v \in A^* \quad ubbmbbv \in X^*$$

Puisque b est une lettre baïonnette, la factorisation de ce mot en mots de X impose que

$$bmb = x_1 x_2 \cdots x_n \quad x_i \in X \quad n > 1$$

Posons $x_1 = bq \quad x_n = pb$. Cette construction associe à tout mot m un unique couple (p, q) tel $m = pxq$ où $x \in X^*$. Notons P et Q les ensembles des mots p et q ainsi obtenus. Montrons que $A^* = QX^*P$ où les produits sont non-ambigus. L'égalité se déduit de la définition même des ensembles P et Q. Pour montrer la non ambiguïté, supposons qu'un même mot se factorise en $qxp = q'x'p'$. Par construction $bqxpb = bq'xp'b \in X^*$ avec $bq, pb, bq', q'b \in X^*$; donc $p = p'$, $q = q'$ et $x = x'$. ∎

Définition 3.4 Étant donné un automate $\mathcal{A}$, une lettre b de l'alphabet est *de rang 1* si (i, j) et (k, l) sont des arcs étiquetés b, alors (i, l) est aussi un arc étiqueté b : l'ensemble des arcs étiquetés b est donc un produit cartésien $c \times l$.

Proposition 3.5 *Un automate de code possédant une lettre de rang 1 reconnaît un code baïonnette.*

Preuve. Soit b la lettre de rang 1 et supposons que le mot $ubvbw$ soit un mot du code. Dans le chemin qu'il détermine de e à e, la première occurrence de b mène de i à j et la deuxième de k à l. Donc, b est aussi étiquette d'un arc de k à j, ce qui crée un cycle (jk) étiqueté vb, évitant e, contrairement à la définition d'un automate de code. ∎

3.2 Prolongement et factorisation

Définition 3.6 Un code X sur l'alphabet A est *prolongeable* en un code Y sur l'alphabet $B \supseteq A$ si $X \subset Y$.

Rappelons que Y est fini et maximal, donc que $X = Y$ si $A = B$.

Proposition 3.7 *Un code prolongeable en un code factorisant est factorisant.*

Preuve. Soit X un code sur A se prolongeant en un code Y sur $B \supseteq A$. Par hypothèse il existe $R, S \subset B^*$ tels que

$$\underline{Y} - 1 = \underline{R}(\underline{B} - 1)\underline{S}$$

Posons $P = R \cap A^*$, $Q = S \cap A^*$ et soit T le polynôme défini par $T - 1 = \underline{P}(\underline{A} - 1)\underline{Q}$. Alors $\underline{Y} - T$ est un polynôme où tous les mots possèdent au moins une occurrence d'une lettre de $B \backslash A$. Il s'en suit que T est le polynôme caractéristique de $Y \cap A^* = X$.

∎

Comme pour les codes, nous dirons qu'un automate de code $\mathcal{A}$ sur A se prolonge en un automate de code $\mathcal{B}$ sur $B \supseteq A$ si la restriction à A de l'automate $\mathcal{B}$ est $\mathcal{A}$. On obtient immédiatement le corollaire :

Corollaire 3.8 *Si un automate de code $\mathcal{A}$ peut se prolonger en un automate de code possédant une lettre de rang 1, alors l'automate $\mathcal{A}$ reconnaît un code factorisant.*

4 Pétales et factorisation

Nous sommes maintenant en mesure d'énoncer le résultat principal.

Théorème 4.1 *Un code est factorisant ssi son automate en pétales se prolonge en un automate de code possédant une lettre de rang 1.*

Preuve. La condition suffisante se déduit du corollaire 3.8. Montrons la réciproque. Soit $b \notin A$ une nouvelle lettre. Considérons maintenant un code factorisant X vérifiant $A^* = QX^*P$. D'après le lemme 1.8, choisissons $p_0 \in P$ et $q_0 \in Q$ de longueurs maximales. Pour tout $p \in P$ et $q \in Q$ il existe des mots u et v tels que $puq_0 \in X$ et $p_0vq \in X$. On ajoute alors à l'automate en pétales des arcs étiquetés b allant de $p|uq_0$ à $p_0v|q$ pour $p \in P$ et $q \in Q$ ce qui construit l'automate $\mathcal{B}$. Cette construction assure l'absence de cycle dans $\mathcal{B}$ car p_0v ne peut être préfixe de p et de même pour les $q \in Q$.

L'automate ainsi obtenu est un automate reconnaissant $Y = X \cup PbQ$. La lettre b étant extérieure à A, on en déduit :

$$\underline{Y} - 1 = \underline{P}(\underline{A} + b - 1)\underline{Q}$$

ce qui prouve que Y est un code (proposition 1.9) et que $\mathcal{B}$ est un automate de code.

■

Le monoïde de transition d'un automate en pétales est un monoïde de relations non-ambigu. On possède des caractérisations combinatoires de ces monoïdes faisant intervenir les boîtes [Boe91]. Ce résultat permet de poser le problème de la factorisation en une question combinatoire sur les boîtes. Plus précisément, dans le cas où l'idéal minimal est de rang 1 (les H-classes sont des singletons), il s'agit de construire une relation de rang 1 (les boîtes permettent de les caractériser) qui donnera la lettre baïonnette. La difficulté est d'assurer que cette relation ne créera pas de cycle dans l'automate.

Une autre question, proposée par un des rapporteurs, amène à faire un lien entre la conjecture de la factorisation et celle de la décidabilité de la complétion finie d'un code. Cette dernière conjecture pose la question de savoir si l'on peut décider si un code fini (non nécessairement maximal) est inclus dans un code maximal fini.

Conjecture 4.2 *Si un code X sur l'alphabet A est prolongeable en un code Y sur l'alphabet $A \cup \{b\}$ (où b est une lettre extérieure à A) alors il est aussi prolongeable en un code baïonnette sur $A \cup \{b\}$.*

On déduirait de ce résultat : Un code X est factorisant *ssi* le code $X \cup \{b\}$ est finiment complétable.

Remerciements

Tous mes remerciements aux rapporteurs qui m'ont aidé dans la rédaction de cet article.

Références

[Boe91] J.-M. Boë, Les boîtes, *Theoret. Comput. Sci.* **81** (1991) 17–34.

[BP85] J. Berstel and D. Perrin, *The Theory of Codes*, Academic Press, New York (1985).

[FR86] C. De Felice and C. Reutenauer, Solution partielle de la conjecture de factorisation des codes, *C.R. Acad. Sci. Paris* **302** (1986) 169–170.

[Rest77] A. Restivo, On codes having no finite completion, *Discrete Math.* **17** (1977) 309–316.

[Reu85] C. Reutenauer, Non-commutative factorization of variable-length codes, *J. Pure Applied Algebra* **36** (1985) 167–186.

[Sch65] M.-P. Schützenberger, On a factorization of free submonoids, *Proc. Amer. Math. Soc.* **36** (1965) 21–24.

Jean-Marie Boë
L.I.R.M.M.
161 rue Ada
F-34392 Montpellier Cedex 5 (France)