

I – Mots qui commutent

On fixe un alphabet Σ . Tous les mots considérés sont des éléments de Σ^* .

1. Soient u et v deux mots non vides. Montrer que les trois conditions suivantes sont équivalentes :

- $uv = vu$
- Il existe un mot w et des entiers $k, p \in \mathbb{N}$ tels que $u = w^k$ et $v = w^p$.
- Il existe $m, n \in \mathbb{N}^*$ tels que $u^m = v^n$.

a) $\Rightarrow$ b). On raisonne par récurrence forte sur $|u| + |v|$. Si $|u| = |v|$, les premières lettres de $uv = vu$ donnent $u = v$. On prend $w = u$ et $k = p = 1$. Cela traite notamment le cas initial $|u| + |v| = 2$.

Sinon, supposons $|u| < |v|$, l'autre cas étant symétrique. Comme $uv = vu$, le mot u est préfixe de v : il existe un mot non vide v' tel que $v = uv'$. Ainsi $u^2v' = uv'u$, puis $uv' = v'u$ par simplification du préfixe u . Comme $|u| + |v'| < |u| + |v|$, l'hypothèse de récurrence donne $u = w^k$ et $v' = w^p$, avec $k, p \geq 1$. Alors $v = w^{k+p}$.

b) $\Rightarrow$ c). Puisque u et v ne sont pas vides, $k, p \geq 1$. Avec $m = p$ et $n = k$, on obtient $u^m = w^{kp} = v^n$.

c) $\Rightarrow$ a). Supposons $u^m = v^n$, avec $m, n \geq 1$. Le mot v est préfixe de u^m , donc uv est préfixe de u^{m+1} , puis de u^{2m} puisque $m+1 \leq 2m$. De même, vu est préfixe de v^{2n} . Comme $u^{2m} = v^{2n}$, les mots uv et vu sont deux préfixes de même taille d'un même mot. Ils sont donc égaux.

Soit $u \in \Sigma^+$. On appelle *commutant* de u le langage $C(u) = \{v \in \Sigma^* \mid uv = vu\}$.

2. Montrer que $C(u)$ est un *sous-monoïde* de $(\Sigma^*, \cdot)$, c'est-à-dire qu'il contient ε et est stable par concaténation.

Il est clair que $\varepsilon \in C(u)$. Si $v, w \in C(u)$, alors $uvw = vuw = vwu$, donc $vw \in C(u)$. Ainsi $C(u)$ est un sous-monoïde de Σ^* .

3. Montrer qu'il existe un unique mot non vide r tel que $C(u) = r^*$.

Comme $u \in C(u)$ et $u \neq \varepsilon$, choisissons un mot non vide $r \in C(u)$ de taille minimale. Toutes ses puissances appartiennent à $C(u)$. Montrons réciproquement, par récurrence forte sur $|v|$, que tout $v \in C(u)$ est une puissance de r . C'est immédiat si $v = \varepsilon$.

Soit $v \in C(u)$ non vide. Par minimalité, $|r| \leq |v|$. Pour tout entier N , on a $u^N r = r u^N$ et $u^N v = v u^N$. En choisissant N tel que $|u^N| \geq |v|$, on voit que r et v sont des préfixes de u^N . Donc r est préfixe de v : écrivons $v = rt$. L'égalité $uv = vu$ donne $urt = rtu$, puis $rut = rtu$ puisque $ur = ru$. Par simplification, $ut = tu$, donc $t \in C(u)$. Comme $|t| < |v|$, l'hypothèse de récurrence donne $t = r^k$, puis $v = r^{k+1}$.

Enfin, si un mot non vide s engendre aussi $C(u)$, il existe $a, b \geq 1$ tels que $r = s^a$ et $s = r^b$. Les tailles donnent $|r| = ab|r|$, donc $a = b = 1$ et $r = s$.

II – Mots de Fibonacci

Les mots de Fibonacci sur l'alphabet $\Sigma = \{a, b\}$ sont définis par

$$f_0 = a, \quad f_1 = b, \quad f_{n+2} = f_{n+1}f_n \quad \text{pour } n \geq 0.$$

1. Montrer que pour $n \geq 2$, le suffixe de taille 2 de f_n est ba si n est pair, ab si n est impair.

On procède par récurrence double. On a $f_2 = ba$ et $f_3 = bab$. Pour $n \geq 4$, le mot f_{n-2} est un suffixe de f_n et a au moins deux lettres. Puisque n et $n-2$ ont la même parité, l'hypothèse de récurrence donne le suffixe annoncé.

2. Pour $n \geq 2$, on note g_n le préfixe de f_n obtenu en supprimant ses deux dernières lettres. Montrer que g_n est un palindrome, c'est-à-dire que $g_n = \widetilde{g_n}$, où $\widetilde{g_n}$ est le mot obtenu en inversant ses lettres.

On procède par récurrence forte. Les mots $g_2 = \varepsilon$, $g_3 = b$ et $g_4 = bab$ sont des palindromes.

Soit $n \geq 5$. Si n est pair, la question précédente donne

$$f_n = f_{n-2}f_{n-3}f_{n-2} = g_{n-2}bag_{n-3}abg_{n-2}ba.$$

Donc $g_n = g_{n-2}bag_{n-3}abg_{n-2}$, qui est un palindrome puisque g_{n-2} et g_{n-3} le sont par hypothèse de récurrence. Si n est impair, on obtient de même

$$g_n = g_{n-2}abg_{n-3}bag_{n-2},$$

qui est aussi un palindrome.

III – Règles sur les expressions régulières

Pour chacune des propositions suivantes sur des expressions régulières quelconques, donner une preuve ou un contre-exemple.

Le symbole $\equiv$ signifie que les expressions décrivent le même langage.

$$1. \quad (e^2)^* \equiv (e^*)^2 \quad (\text{où } e^2 = ee).$$

| Faux avec $e = a$: le mot a n'appartient pas à $\mathcal{L}((aa)^*)$, mais $a = a\varepsilon$ appartient à $\mathcal{L}((a^*)^2)$.

$$2. \quad (e_1 \mid e_2)^* \equiv e_1^* \mid e_2^*.$$

| Faux avec $e_1 = a$ et $e_2 = b$: $ab \in \mathcal{L}((a|b)^*)$, mais $ab \notin \mathcal{L}(a^*|b^*)$.

$$3. \quad (e_1 e_2)^* \equiv e_1^* e_2^*.$$

| Faux avec $e_1 = a$ et $e_2 = b$: $abab \in \mathcal{L}((ab)^*)$, mais $abab \notin \mathcal{L}(a^*b^*)$.

$$4. \quad (e_1 \mid e_2)^* \equiv (e_1^* e_2^*)^*.$$

| Vrai. Posons $A = \mathcal{L}(e_1)$ et $B = \mathcal{L}(e_2)$. Un mot de $(A^*B^*)^*$ est une concaténation de mots de A et de B , donc appartient à $(A \cup B)^*$. Réciproquement, $A \cup B \subseteq A^*B^*$ puisque chaque étoile contient ε . L'étoile conservant l'inclusion, $(A \cup B)^* \subseteq (A^*B^*)^*$.

IV – Exemples de langages réguliers

1. Donner une expression régulière dont le langage est l'ensemble des mots sur $\{a, b, c\}$ contenant exactement un a et un b (et un nombre quelconque de c).

| En distinguant l'ordre d'apparition de a et b , on obtient

$$c^*ac^*bc^* \mid c^*bc^*ac^*.$$

2. Donner une expression régulière dont le langage est l'ensemble des mots sur $\{a, b, c\}$ ne contenant pas de a consécutifs (aa ne doit pas apparaître).

| L'expression $(a(b|c)|b|c)^*(a|\varepsilon)$ convient : un a qui n'est pas en dernière position doit être suivi d'un b ou d'un c .

3. Donner une expression régulière dont le langage est l'ensemble des mots sur $\{a, b, c\}$ contenant exactement deux a et tels que tout c est immédiatement précédé d'un b .

| Posons $e = (b|bc)^*$, qui décrit les mots sur $\{b, c\}$ dont chaque c est immédiatement précédé d'un b . L'expression $eaeae$ convient.

Si $x \in \mathbb{R}$, on note $L(x)$ l'ensemble des préfixes de la suite des chiffres de x après la virgule. On utilise le développement décimal qui ne se termine pas par une infinité de 9. Par exemple,

$$L(\pi) = \{\varepsilon, 1, 14, 141, 1415, \dots\}.$$

4. En sachant que $\frac{1}{6} = 0.1666\dots$ et $\frac{1}{7} = 0.142857142857\dots$, montrer que $L(\frac{1}{6})$ et $L(\frac{1}{7})$ sont réguliers.

| Une expression pour $L(\frac{1}{6})$ est $\varepsilon \mid 16^*$. Une expression pour $L(\frac{1}{7})$ est

$$(142857)^*(\varepsilon \mid 1 \mid 14 \mid 142 \mid 1428 \mid 14285 \mid 142857).$$

5. Montrer plus généralement que $L(x)$ est régulier si $x \in \mathbb{Q}$ (on montrera plus tard que c'est en fait une équivalence).

| Le développement décimal d'un rationnel est ultimement périodique : la partie après la virgule s'écrit $tppp\dots$, avec t un mot fini et p un mot fini non vide (on peut prendre $p = 0$ pour un décimal). Notons $\text{Pref}(m)$ l'ensemble fini des préfixes de m . On a

$$L(x) = \text{Pref}(t) \cup \{t\}\{p\}^* \text{Pref}(p).$$

| Un préfixe s'arrête soit dans t , soit après t , un certain nombre de copies de p et un préfixe de p . Ce langage est régulier par stabilité par union, concaténation et étoile.

6. Donner une expression régulière dont le langage est $\{a^n b^p \mid n, p \in \mathbb{N}, n \equiv p \pmod{2}\}$.

| Les deux exposants sont pairs ou tous deux impairs. L'expression

$$(aa)^*(bb)^* \mid a(aa)^*b(bb)^*$$

| convient.

V – Distance de Hamming

On fixe un alphabet fini Σ . Si $u = u_1\dots u_n$ et $v = v_1\dots v_n$ sont deux mots de même taille, leur distance de Hamming est

$$d(u, v) = |\{i \in \{1, \dots, n\} \mid u_i \neq v_i\}|.$$

1. Montrer que, pour tout $n \in \mathbb{N}$, la distance de Hamming est une distance sur Σ^n .

La distance est positive ou nulle, symétrique, et $d(u, v) = 0$ si et seulement si $u = v$. Pour l'inégalité triangulaire, soient $u, v, w \in \Sigma^n$. Si $u_i \neq w_i$, alors $u_i \neq v_i$ ou $v_i \neq w_i$. Ainsi,

$$\{i \mid u_i \neq w_i\} \subseteq \{i \mid u_i \neq v_i\} \cup \{i \mid v_i \neq w_i\}.$$

En prenant les cardinaux, on obtient $d(u, w) \leq d(u, v) + d(v, w)$.

Étant donné un langage L sur Σ , on définit son voisinage de Hamming par

$$\mathcal{H}(L) = \{u \in \Sigma^* \mid \exists v \in L, |u| = |v| \text{ et } d(u, v) \leq 1\}.$$

2. Sur l'alphabet $\{0, 1\}$, donner une expression régulière pour $\mathcal{H}(\mathcal{L}(0^*1^*))$.

On peut conserver le mot, changer un 0 en 1, ou changer un 1 en 0. On obtient l'expression

$$0^*1^* \mid 0^*10^*1^* \mid 0^*1^*01^*.$$

3. Montrer que si L est un langage régulier, alors $\mathcal{H}(L)$ est régulier.

On raisonne par induction structurelle sur les expressions régulières. Si L est fini, $\mathcal{H}(L)$ est fini : pour un mot m , il y a $1 + |m|(|\Sigma| - 1)$ mots à distance au plus 1 de m .

Pour l'union et la concaténation, on utilise les identités

$$\mathcal{H}(L_1 \cup L_2) = \mathcal{H}(L_1) \cup \mathcal{H}(L_2),$$

$$\mathcal{H}(L_1 L_2) = \mathcal{H}(L_1) L_2 \cup L_1 \mathcal{H}(L_2).$$

Une modification d'au plus une lettre d'un mot $u_1 u_2$ se situe dans l'un des deux facteurs. Les membres de droite sont réguliers par hypothèse d'induction et stabilité par union et concaténation.

Enfin,

$$\mathcal{H}(L^*) = \{\varepsilon\} \cup L^* \mathcal{H}(L) L^*.$$

Le mot vide est toujours présent. Pour une concaténation non vide de facteurs de L , on peut modifier au plus une lettre dans l'un d'entre eux, ou ne rien modifier. Cette identité reste vraie pour $L = \emptyset$ et achève l'induction.

4. Sur l'alphabet $\Sigma = \{0, 1\}$, écrire une fonction `f : int regexp -> int regexp` renvoyant une expression régulière du voisinage de Hamming d'un langage, en utilisant le type suivant. Les lettres de l'expression d'entrée valent 0 ou 1.

```
type 'a regexp =
| Vide | Epsilon | L of 'a (* L a est la lettre a *)
| Union of 'a regexp * 'a regexp
| Concat of 'a regexp * 'a regexp
| Etoile of 'a regexp
```

On traduit les identités de la question précédente :

```
let rec f = function
| Vide -> Vide
| Epsilon -> Epsilon
| L _ -> Union (L 0, L 1)
| Union (e1, e2) -> Union (f e1, f e2)
| Concat (e1, e2) ->
  Union (Concat (f e1, e2), Concat (e1, f e2))
| Etoile e ->
  Union (Epsilon, Concat (Etoile e, Concat (f e, Etoile e)))
```

L'union avec Epsilon est nécessaire, notamment pour Etoile Vide.

VI – Clôture par sur-mot

ENS MP oral

On fixe un alphabet fini non vide Σ . Étant donnés deux mots $w, w' \in \Sigma^*$, on dit que w' est un sur-mot de w , noté $w \preceq w'$, s'il existe une fonction strictement croissante φ de $\{1, \dots, |w|\}$ dans $\{1, \dots, |w'|\}$ telle que $w_i = w'_{\varphi(i)}$ pour tout $1 \leq i \leq |w|$. Étant donné un langage L , on note $\overline{L}$ le langage des sur-mots de mots de L :

$$\overline{L} = \{w' \in \Sigma^* \mid \exists w \in L, w \preceq w'\}.$$

1. On pose $L_0 = \mathcal{L}(ab^*a)$ et $L_1 = \mathcal{L}((ab)^*)$, sur un alphabet contenant a et b . Donner une expression régulière pour $\overline{L_0}$ et $\overline{L_1}$.

$\overline{L_0}$ est le langage des mots contenant au moins deux a , soit $\Sigma^* a \Sigma^* a \Sigma^*$. En effet, tout sur-mot d'un mot de L_0 contient deux a et, réciproquement, un mot contenant deux a est un sur-mot de $aa \in L_0$. On a $\overline{L_1} = \Sigma^*$ puisque $\varepsilon \in L_1$.

2. Montrer que, pour tout langage L , on a $\overline{\overline{L}} = \overline{L}$.

La relation $\preceq$ est réflexive et transitive : pour la transitivité, on compose les fonctions strictement croissantes témoignant de $w \preceq w'$ et $w' \preceq w''$. Leur composition témoigne de $w \preceq w''$.

Par réflexivité, $\bar{L} \subseteq \bar{\bar{L}}$. Réciproquement, si $u'' \in \bar{\bar{L}}$, il existe $u' \in \bar{L}$ tel que $u' \preceq u''$, puis $u \in L$ tel que $u \preceq u'$. Par transitivité, $u \preceq u''$, donc $u'' \in \bar{L}$.

3. Existe-t-il des langages L' pour lesquels il n'existe aucun langage L tel que $\bar{L} = L'$?

Oui : tout langage fini non vide convient. Si L est non vide et $m \in L$, alors $m\Sigma^* \subseteq \bar{L}$, donc $\bar{L}$ est infini puisque l'alphabet est non vide. Si $L = \emptyset$, sa clôture est vide.

On peut aussi prendre $L' = L_0$: si $\bar{L} = L_0$, l'idempotence donnerait $L_0 = \bar{L_0}$, ce qui est faux d'après la question 1.

4. Montrer que, pour tout langage régulier L , le langage $\bar{L}$ est régulier.

Soit A un automate fini non déterministe reconnaissant L . On construit A' en ajoutant, à chaque état, une boucle pour chaque lettre de Σ . Les états initiaux et finaux restent inchangés.

Si u est accepté par A et $u \preceq u'$, un chemin acceptant pour u' suit les transitions de A aux positions correspondant aux lettres de u , et les boucles ajoutées aux autres positions. Réciproquement, d'un chemin acceptant dans A' , on retire les passages par les boucles ajoutées : les transitions restantes forment un chemin acceptant dans A , portant un sous-mot du mot initial. Ainsi $\mathcal{L}(A') = \bar{L}$.

Une autre preuve procède par induction structurelle avec les identités

$$\begin{aligned}\bar{\emptyset} &= \emptyset, \quad \overline{\{\varepsilon\}} = \Sigma^*, \quad \overline{\{a\}} = \Sigma^* a \Sigma^*, \\ \overline{L_1 L_2} &= \overline{L_1} \overline{L_2}, \quad \overline{L_1 \cup L_2} = \overline{L_1} \cup \overline{L_2}, \\ \overline{L^*} &= \Sigma^*.\end{aligned}$$

La dernière identité vient de $\varepsilon \in L^*$; en revanche, $\bar{L^*} = \bar{L}^*$ est faux en général (prendre $\Sigma = \{a, b\}$ et $L = \{a\}$).

5. On admet pour cette question le résultat suivant : pour toute suite $(w_n)_{n \in \mathbb{N}}$ de mots de Σ^* , il existe $i < j$ tels que $w_i \preceq w_j$.

Montrer que, pour tout langage L , non nécessairement régulier, il existe un langage fini $F \subseteq L$ tel que $\bar{F} = \bar{L}$.

Si $L = \emptyset$, prendre $F = \emptyset$. Sinon, on énumère les mots de L dans une suite infinie (w_n) , éventuellement avec des répétitions. Une position i est dite innovante s'il n'existe aucun $j < i$ avec $w_j \preceq w_i$. On prend $F = \{w_i \mid i \text{ est innovante}\}$.

Il n'y a qu'un nombre fini de positions innovantes : sinon, la sous-suite des mots situés à ces positions contredirait le résultat admis. Ainsi F est fini, et $F \subseteq L$ donne $\bar{F} \subseteq \bar{L}$.

Par récurrence sur i , on montre que $w_i \in \bar{F}$. Si i est innovante, $w_i \in F$. Sinon, il existe $j < i$ tel que $w_j \preceq w_i$; l'hypothèse de récurrence et la transitivité donnent encore $w_i \in \bar{F}$. Ainsi $L \subseteq \bar{F}$ et, par monotonie et par la question 2, $\bar{L} \subseteq \bar{\bar{F}} = \bar{F}$.

6. Un langage L est clos par sur-mots si, pour tout $u \in L$ et tout $v \in \Sigma^*$ tel que $u \preceq v$, on a $v \in L$. Dédurre de la question précédente que tout langage clos par sur-mots est régulier.

Par définition, $L = \bar{L}$. La question 5 fournit un langage fini $F \subseteq L$ tel que $\bar{F} = \bar{L} = L$. Le langage F est régulier car fini ; sa clôture est régulière par la question 4.

7. On admet que les langages réguliers sont stables par complémentaire. Un langage L est clos par sous-mots si, pour tout $u \in L$ et tout $v \in \Sigma^*$ tel que $v \preceq u$, on a $v \in L$. Montrer que tout langage clos par sous-mots est régulier.

Son complémentaire $L' = \Sigma^* \setminus L$ est clos par sur-mots. En effet, si $u \in L'$ et $u \preceq v$, on ne peut avoir $v \in L$, car la clôture par sous-mots donnerait $u \in L$, contradiction. Par la question 6, L' est régulier. Son complémentaire L est donc régulier.

8. Démontrer le résultat admis à la question 5.

C'est le lemme de Higman pour un alphabet fini. Appelons mauvaise une suite (w_n) pour laquelle il n'existe aucun $i < j$ tel que $w_i \preceq w_j$. Supposons qu'une telle suite existe.

Construisons une mauvaise suite minimale : w_0 est de taille minimale parmi les mots pouvant commencer une mauvaise suite ; une fois $w_0, \dots, w_{n-1}$ fixés, on choisit w_n de taille minimale parmi les mots permettant de prolonger ce préfixe en une mauvaise suite. La suite obtenue est mauvaise, car chaque préfixe possède un prolongement mauvais.

Aucun w_n n'est vide, puisque ε est sous-mot de tout mot suivant. Comme Σ est fini, une lettre a commence une infinité de mots de la suite. Écrivons $w_{i_j} = av_j$, où $i_0 < i_1 < \dots$ sont les indices correspondants. Considérons la nouvelle suite

$$w_0, \dots, w_{i_0-1}, v_0, v_1, v_2, \dots$$

Elle est encore mauvaise :

- deux mots du préfixe ne sont pas comparables dans cet ordre, par hypothèse ;

- si $w_r \preccurlyeq v_j$ avec $r < i_0$, alors $w_r \preccurlyeq w_{i_j}$, contradiction ;
- si $v_j \preccurlyeq v_k$ avec $j < k$, alors $av_j \preccurlyeq av_k$, donc $w_{i_j} \preccurlyeq w_{i_k}$, contradiction.

Cette suite coïncide avec la suite minimale jusqu'à l'indice $i_0 - 1$, puis contient v_0 , plus court d'une lettre que w_{i_0} . Elle contredit le choix minimal de w_{i_0} . Il n'existe donc pas de mauvaise suite.