I Mots qui commutent

Soient u et v deux mots. Montrer que les deux conditions suivantes sont équivalentes :

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

Solution : $2 \Rightarrow 1$ est évident. On montre $1 \Rightarrow 2$ par récurrence forte sur $n = |u| + |v|$.
Cas de base : $|u| + |v| = 0$. Si $u = v = \varepsilon$, alors $u = w^1$ et $v = w^1$ pour $w = \varepsilon$, $n = p = 1$.
Cas inductif : Soit $n \in \mathbb{N}^*$. Supposons que la propriété est vraie pour des mots u, v tels que $|u| + |v| < n$.
Soient u et v tels que $uv = vu$ et $|u| + |v| < n$.
Si $|u| = |v|$ alors les $|u|$ premières lettres dans l'égalité $uv = vu$ donne $u = v$ et $u = w^1 = v$ avec $w = u$.
Supposons $|u| \leq |v|$ (l'autre cas étant symétrique). Comme $uv = vu$, u est préfixe de v : il existe un mot $v' \neq \varepsilon$ tel que $v = uv'$. On a alors $u^2v' = uv'u$. En particulier, $uv' = v'u$. Comme $|u| + |v'| < |u| + |v|$, il existe un mot w et des entiers $k, p \geq 1$ tels que $u = w^k$ et $v' = w^p$, par hypothèse de récurrence. On a alors $u = w^k$ et $v = uv' = w^{k+p}$, ce qui conclut la preuve.

II Mots de Fibonacci

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

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

1. Montrer que pour $n \geq 2$, le suffixe de longueur 2 de f_n est ba si n est pair, ab si n est impair.
2. Pour $n \geq 3$, 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 obtenu en inversant les lettres de g_n.

Solution :

1. On procède par récurrence double sur $n \geq 2$.
   - On a $f_2 = ba$ et $f_3 = bab$ donc la propriété est vérifiée pour $n = 2$ et $n = 3$.
   - Si $n \geq 4$ est pair, alors f_{n-2} est un suffixe de f_n et $n-2$ est pair, donc par hypothèse de récurrence ba est suffixe de f_{n-2} et donc de f_n. De même dans le cas où $n \geq 4$ est impair.
2. On procède à nouveau par récurrence.
   - $g_2 = \varepsilon$, $g_3 = b$ et $g_4 = bab$ qui sont bien des palindromes.
   - Soit $n \geq 5$, supposons la propriété vérifiée jusqu'à $n - 1$. Si n est pair, on a $f_n = f_{n-1}f_{n-2} = f_{n-2}f_{n-3}f_{n-2} = g_{n-2}bag_{n-3}abg_{n-2}ba$ donc $g_n = g_{n-2}bag_{n-3}abg_{n-2}$. Comme g_{n-2} et g_{n-3} sont des palindromes par hypothèse de récurrence, g_n en est également un. On procède de même dans le cas n impair.

III Règles sur les expressions régulières

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

1. $(e^*)^* \equiv e^*$
2. $(e_1|e_2)^* \equiv e_1^*|e_2^*$
3. $(e_1e_2)^* \equiv e_1^*e_2^*$
4. $(e_1|e_2)^* \equiv (e_1^*e_2^*)^*$

Solution :

1. Vrai. Voir cours.
2. Faux car $ab \in (a|b)^*$ mais $ab \notin a^* + b^*$.
3. Faux car $abab \in (ab)^*$ mais $abab \notin a^*b^*$.
4. Vrai. Voir cours.

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).

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).
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 précédé d'un b.
4. Si $x \in \mathbb{R}$, on note $L(x)$ l'ensemble des préfixes des chiffres de x après la virgule. Par exemple, $L(\pi) = \{\varepsilon, 1, 14, 141, 1415...\}$. En sachant que $\frac{1}{6} = 0.1666...$ et $\frac{1}{7} = 0.142857142857...$, montrer que $L(\frac{1}{6})$ et $L(\frac{1}{7})$ sont réguliers.
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).
6. Donner une expression régulière dont le langage est $\{a^n b^p \mid n = p \mod 2\}$.

Solution :

1. En distinguant le cas où a est avant b et le cas où b est avant a: $c^*ac^*bc^*|c^*bc^*ac^*$.
2. On peut donner $(a(b|c)|b|c)^*(a|\varepsilon)$ (un a doit être suivi d'un b ou d'un c).
3. Soit $e = (b|bc)^*$ (décrivant tous les mots sur $\{b, c\}$ dont chaque c est précédé d'un b). Alors $eaeae$ est une expression régulière qui convient.
4. $\varepsilon|16^*$ est une expression régulière de langage $L(\frac{1}{6})$.

   $(142857)^*(\varepsilon|1|14|142|1428|14285|142857)$ est une expression régulière de langage $L(\frac{1}{7})$.
5. Si $x \in \mathbb{Q}$, on peut écrire ses chiffres sous la forme $x = x_1, x_2ppp...$
   Soit $Pref(m)$ l'ensemble des préfixes d'un mot m, qui est un ensemble fini si m est fini ($|Pref(m)| = |m| + 1$). Alors $L(x) = Pref(x_2)|x_2p^*Pref(p)$ (un élément de $L(x)$ est soit un préfixe de x_2 soit contient x_2 suivi d'un certain nombre de p, suivi d'une partie de p).
6. Soit il y a un nombre pair de a et un nombre pair de b, soit il y a un nombre impair de a et un nombre impair de b : $(aa)^*(bb)^*|a(aa)^*b(bb)^*$.

V Distance de Hamming

Si $u = u_1...u_n$ et $v = v_1...v_n$ sont deux mots de même longueur sur un alphabet Σ, leur distance de Hamming est :

$$d(u, v) = |\{i \mid u_i \neq v_i\}|$$

1. Montrer que la distance de Hamming est une distance sur Σ^*.
   ► Soient $u = u_1...u_n, v = v_1...v_n, w = w_1...w_n$ trois mots de même taille. Si $u_i \neq w_i$ alors $u_i \neq v_i$ ou $v_i \neq w_i$ (sinon, $u_i = v_i = w_i$). D'où $d(u, v) + d(v, w) \leq d(u, w)$. $d(u, v) = d(v, u)$ et $d(u, v) = 0 \Leftrightarrow u = v$ sont facilement vérifiés.

Étant donné un langage L sur Σ, on définit son voisinage de Hamming $\mathcal{H}(L) = \{u \in \Sigma^* \mid \exists v \in L,\ d(u, v) \leq 1\}$.

2. Donner une expression régulière pour $\mathcal{H}(L(0^*1^*))$.
   ► C'est l'ensemble des mots obtenus en changeant un 0 par un 1 ou inversement, c'est à dire $L(0^*10^*1^*|0^*1^*01^*)$.
3. Montrer que si L est un langage régulier alors $\mathcal{H}(L)$ est un langage régulier.
   ►
   - $f(0) = 1,\ f(1) = 0$
   - $f(e_1e_2) = f(e_1)e_2|e_1H(e_2)$: modifier une lettre de $u = u_1u_2 \in L(e_1e_2)$ revient à modifier une lettre de u_1 ou un lettre de u_2.
   - $f(e_1|e_2) = f(e_1)|f(e_2)$.
   - Si $e = e_1^*$: $f(e_1^*) = e_1^*f(e_1)e_1^*$.
4. Écrire une fonction `f : 'a regexp -> 'a regexp` renvoyant une expression régulière pour le voisinage de Hamming d'un langage, en utilisant le type suivant :

```ocaml
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
```

►

VI Clôture par sur-mot (oral ENS info)

On fixe un alphabet Σ. Étant donné deux mots $w, w' \in \Sigma^*$, on dit que w' est un sur-mot de w, noté $w \preccurlyeq w'$, s'il existe une fonction strictement croissante ϕ de $\{1, \ldots, |w|\}$ dans $\{1, \ldots, |w'|\}$ telle que $w_i = w'_{\phi(i)}$ pour tout $1 \leq i \leq |w|$, où $|w|$ dénote la longueur de w et w_i dénote la i-ème lettre de w. Étant donné un langage L, on note $\overline{L}$ le langage des sur-mots de mots de L, c'est-à-dire $\overline{L} := \{w' \in \Sigma^* \mid \exists w \in L, w \preccurlyeq w'\}$.

1. On pose L_0 le langage défini par l'expression régulière ab^*a, et L_1 le langage défini par l'expression régulière $(ab)^*$. Donner une expression régulière pour $\overline{L_0}$ et pour $\overline{L_1}$.
2. Montrer que, pour tout langage L, on a $\overline{\overline{L}} = \overline{L}$.
3. Existe-t-il des langages L' pour lesquels il n'existe aucun langage L tel que $\overline{L} = L'$?
4. Montrer que, pour tout langage régulier L, le langage $\overline{L}$ est également régulier.
5. On admettra 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 $w_i \preccurlyeq w_j$.

   Montrer que, pour tout langage L (non nécessairement régulier), il existe un langage fini $F \subseteq L$ tel que $\overline{F} = \overline{L}$.
6. Un langage L est clos par sur-mots si, pour tout $u \in L$ et $v \in \Sigma^*$ tel que $u \preccurlyeq v$, on a $v \in L$. Déduire de la question précédente que tout langage clos par sur-mots est régulier.
7. On admet que les langages réguliers sont stables par passage au complémentaire. Un langage L est clos par sous-mots si, pour tout $u \in L$ et $v \in \Sigma^*$ tel que $v \preccurlyeq u$, on a $v \in L$. Montrer que tout langage clos par sous-mots est régulier.
8. Démontrer le résultat admis à la question 5.

Solution :

1. Le langage $\overline{L_0}$ est le langage des mots qui contiennent deux a, c'est-à-dire $\Sigma^* a \Sigma^* a \Sigma^*$. En effet, tout sur-mot d'un mot de ab^*a doit clairement contenir deux a. Réciproquement, tout mot contenant deux a est un sur-mot de aa qui appartient à L_0.

   Le langage $\overline{L_1}$ est Σ^*, puisque tout mot est un sur-mot de $\varepsilon \in L_1$.
2. On observe d'abord que la relation $\preccurlyeq$ est transitive. En effet, pour tous mots $w, w', w'' \in \Sigma^*$ tels que $w \preccurlyeq w'$ et $w' \preccurlyeq$, en notant ϕ et ϕ' les fonctions strictement croissantes qui en témoignent, leur composition $\phi' \circ \phi$ est une fonction strictement croissante de $\{1, \ldots, |w|\}$ dans $\{1, \ldots, |w''|\}$, et pour tout $1 \leq i \leq w$ on a $w''_{\phi'(\phi(i))} = w'_{\phi(i)} = w_i$.

   On montre à présent l'égalité demandée. Il est clair que $\overline{L} \subseteq \overline{\overline{L}}$, donc on montre l'inclusion inverse. Soit $u'' \in \overline{\overline{L}}$, il existe un mot $u' \in \overline{L}$ tel que $u' \preccurlyeq u$. Par définition de $\overline{L}$, il existe un mot $u \in L$ tel que $u \preccurlyeq u'$. Par transitivité, on a $u \preccurlyeq u''$. Ainsi, on a bien $u'' \in \overline{L}$, ce qui conclut.
3. Pour tout langage non-vide L, le langage $\overline{L}$ est nécessairement infini : en effet, pour $u \in L$ quelconque, on a $u\Sigma^* \subseteq \overline{L}$. Par ailleurs, on a clairement $\overline{\emptyset} = \emptyset$. Ainsi, si l'on prend L' fini non-vide, on sait qu'il n'existe aucun langage L tel que $\overline{L} = L'$.

   Autre preuve possible : on considère le langage L_0. Supposons par l'absurde qu'il existe un langage L tel que $\overline{L} = L_0$. Dans ce cas, on a $\overline{\overline{L}} = \overline{L_0}$, donc d'après la question 1 , on a $\overline{L} = \overline{L_0}$. C'est absurde car L_0 et $\overline{L_0}$ sont manifestement différents. Ainsi, $L' := L_0$ convient.
4. Soit A un automate fini non-déterministe qui reconnaisse le langage régulier L. Construisons un automate A' en ajoutant à chaque état de A une boucle pour toutes les lettres de l'alphabet : formellement, on initialise $A' := A$ et pour chaque $a \in \Sigma$ et chaque état q de A, on ajoute à A' une transition de q à q étiquetée par a.

   Il est clair que, pour tout mot u accepté par A et pour tout mot u' tel que $u \preccurlyeq u'$, le mot u' est accepté par A' :

pour ϕ une fonction strictement croissante qui témoigne du fait que $u \preccurlyeq u'$, il suffit de suivre le chemin pour u dans A' pour les positions de u' appartenant à l'image de ϕ, et de suivre les nouvelles transitions pour les positions de u' qui n'appartiennent pas à l'image de ϕ. Réciproquement, si l'on considère un mot u' accepté par A' et un chemin qui en témoigne, on peut construire un mot u accepté par A tel que $u \preccurlyeq u'$ en considérant la restriction de ce chemin aux transitions de A.

On peut aussi démontrer cette question par induction structurelle sur les expressions régulières à l'aide des identités suivantes :

$-\overline{\emptyset} = \emptyset$

$-\overline{\varepsilon} = \Sigma^*$

$-\overline{a} = \Sigma^* a \Sigma^*$ pour tout $a \in \Sigma$

$-\overline{L_1 L_2} = \overline{L_1}\,\overline{L_2}$

$-\overline{L_1 \cup L_2} = \overline{L_1} \cup \overline{L_2}$

$-\overline{L^*} = \Sigma^*$

La dernière égalité est due au fait que L^* contient toujours le mot vide; en revanche il n'est pas vrai que $\overline{L^*} = \overline{L}^*$, prendre par exemple $L = a$.

5. Soit L un langage quelconque. Si L est vide, on peut prendre $F = \emptyset$ et conclure. Sinon, posons $(w_n)_{n\in\mathbb{N}}$ une suite infinie énumérant les mots du langage L (éventuellement avec des doublons). Une position $i \in \mathbb{N}$ est dite innovante s'il n'existe aucun $j < i$ tel que $w_j \preccurlyeq w_i$. On choisit pour F le sous-ensemble de L formé des mots aux positions innovantes, c'est-à-dire $\{w_i \mid i$ est innovante $\}$.

   On observe à présent qu'il y a un nombre fini de positions innovantes. En effet, dans le cas contraire, la suite extraite obtenue à partir de $(w_n)_{n\in\mathbb{N}}$ en conservant les lettres aux positions innovantes serait un contre-exemple à la question 4 . Ainsi, F est-il bien fini.

   Montrons à présent que $\overline{F} = \overline{L}$. En effet, comme $F \subseteq L$, on a $\overline{F} \subseteq \overline{L}$ par monotonie de la clôture par sur-mots. Pour la réciproque, il suffit de montrer que $L \subseteq \overline{F}$, car cela implique (à nouveau par monotonie de la clôture par sur-mots) que $\overline{L} \subseteq \overline{\overline{F}}$, ce qui implique par la question 1 que $\overline{L} \subseteq \overline{F}$. Montrons par induction sur $i \in \mathbb{N}$ que $w_j \in \overline{F}$ pour tout $j < i$. Le cas de base est tautologique. Pour le cas de récurrence, choisissons $i \in \mathbb{N}$. Soit i est innovante, soit i n'est pas innovante. Dans le premier cas, on a $w_i \in F$ donc $w_i \in \overline{F}$. Dans le second cas, il existe $j < i$ tel que $w_j \preccurlyeq w_i$, et par hypothèse de récurrence on a $w_j \in \overline{F}$, ainsi on a $w_i \in \overline{F}$. Ainsi, dans les deux cas on a $w_i \in \overline{F}$. On a donc établi notre résultat par récurrence, et on a donc bien l'inclusion réciproque $L \subseteq \overline{F}$.

6. Soit L un langage clos par sur-mots. On sait par la question 5 qu'il existe un langage fini $F \subseteq L$ tel que $\overline{F} = \overline{L}$. Or on a $\overline{L} = L$. En effet, il est clair que $L \subseteq \overline{L}$, et réciproquement, pour tout $v \in \overline{L}$, il existe par définition de $\overline{L}$ un mot $u \in L$ tel que $u \preccurlyeq v$, et ainsi $v \in L$ car L est clos par sur-mots. On sait donc que $L = \overline{F}$, et on sait que F est régulier (car fini), donc $\overline{F}$ est régulier par la question 3, ainsi L est-il régulier.

7. On a besoin de pouvoir énumérer efficacement des mots de L qui ne sont pas des sur-mots de mots précédemment énumérés. En d'autres termes, il nous faudrait un oracle qui, étant donné une liste de mots $W = (w_0, \ldots, w_n)$ de L, produit un mot w_{n+1} de L tel que $w_i \not\preccurlyeq w_{n+1}$ pour tout $0 \leq i \leq n$, ou bien conclut qu'aucun tel mot n'existe, de sorte que $\overline{L} = \overline{W}$. On pourrait aussi se contenter d'un oracle qui, étant donné un langage régulier L' (ici, $L' := \overline{W}$), produit un mot de $L \backslash L'$, ou conclut que $L \subseteq L'$

   Étant donné un tel oracle, on construirait petit à petit un automate de $\overline{W}$ jusqu'à avoir couvert tout $\overline{L}$

   On ne peut pas espérer que cette procédure soit efficace, car le nombre d'invocations de l'oracle (et la taille de l'automate construit) serait très grande même pour des langages simples; formellement, même quand L est régulier, elle peut être exponentielle en la longueur d'une expression régulière pour L. Si l'on prend par exemple L_k l'ensemble régulier des mots de longueur k pour un certain $k \in \mathbb{N}$, que l'on peut décrire avec une expression régulière ou un automate de taille $O(k \times |\Sigma|)$, les mots de L_k sont tous incomparables pour la relation $\preccurlyeq$ vu qu'ils sont de même longueur, et il y a un nombre exponentiel de tels mots, à savoir $|\Sigma|^k$

8. Soit L un langage clos par sous-mots, et soit $L' := \Sigma^* \backslash L$ son complémentaire. Montrons que L' est clos par sur-mots. En effet, soit $u \in L'$ et $v \in \Sigma^*$ tels que $u \preccurlyeq v$. Procédons par l'absurde et supposons que $v \notin L'$. On a alors $v \in L$. Comme $u \preccurlyeq v$ et que L est clos par sous-mots, on sait que $u \in L$, et ainsi $u \notin L'$, contredisant notre hypothèse. Ainsi, $v \in L'$, ce qui établit que L' est clos par sur-mots. On sait donc que L' est régulier. Comme les langages réguliers sont clos par complémentation, le complémentaire L de L' est lui aussi régulier.

9. Il s'agit du lemme de Higman dans le cas particulier des alphabets finis.

   Procédons par l'absurde et supposons qu'il existe une mauvaise suite, c'est-à-dire une suite $(w_n)_{n\in\mathbb{N}}$ telle qu'il n'existe pas de $i < j$ telle que $w_i \preccurlyeq w_j$. Construisons une nouvelle suite $(w'_n)_{n\in\mathbb{N}}$ de la façon suivante : le mot w'_0 est un mot de longueur minimale telle qu'il existe une mauvaise suite commençant par w'_0 (un tel w'_0 existe par notre hypothèse), le mot w'_1 est un mot de longueur minimale telle qu'il existe une mauvaise suite commençant par w'_0, w'_1 (un tel w'_1 existe par définition de w'_0), et ainsi de suite. La suite $(w'_n)_{n\in\mathbb{N}}$ ainsi définie est clairement mauvaise : pour tout $i < j$, la définition de w'_j assure qu'on ne peut avoir $w'_i \preccurlyeq w'_j$. Par ailleurs, la définition de $(w'_n)_{n\in\mathbb{N}}$ assure qu'elle est minimale, c'est-à-dire que pour toute mauvaise suite $(w''_n)_{n\in\mathbb{N}}$, si on pose $i \in \mathbb{N}$ le premier indice tel que $w'_i \neq w''_i$, on a nécessairement $|w'_i| \leq |w''_i|$.

   On va aboutir à notre contradiction en construisant à partir de $(w'_n)_{n\in\mathbb{N}}$ une nouvelle mauvaise suite qui contredise sa minimalité. Soit $a \in \Sigma$ une lettre quelconque telle qu'il existe un nombre infini de mots de $(w'_n)_{n\in\mathbb{N}}$ ayant a comme première lettre : comme Σ est fini, un tel a existe nécessairement. Soit $p \in \mathbb{N}$ le plus petit entier tel que w'_p commence par a. On construit la suite $(w''_n)_{n\in\mathbb{N}}$ comme la concaténation de $w'_0, \ldots, w'_{p-1}$ et de la suite extraite de $(w'_n)_{n\in\mathbb{N}}$ des mots commençant par a à qui on a retiré leur première lettre.

   La suite $(w''_n)_{n\in\mathbb{N}}$ est une mauvaise suite. En effet, soit $p < q$. Si $q < i$, alors $w''_p = w'_p$ et $w''_q = w'_q$, donc $w''_p \not\preccurlyeq w''_q$ car $(w'_n)_{n\in\mathbb{N}}$ est mauvaise. Si $i \leq p$, alors $aw''_p = w'_p$ et $aw''_q = w'_q$, donc on conclut encore car $(w'_n)_{n\in\mathbb{N}}$ est mauvaise. Si $p < i \leq q$, alors $w''_p = w'_p$ et $aw''_q = w'_q$, et on conclut de même.

   Par ailleurs, la suite $(w''_n)_{n\in\mathbb{N}}$ contredit la minimalité de $(w'_n)_{n\in\mathbb{N}}$. En effet, le premier indice où ces deux suites diffèrent est p, et on a $\left|w''_p\right| = \left|w'_p\right| - 1$, ce qui contredit bien la minimalité. C'est absurde, et ainsi notre hypothèse initiale affirmant l'existence d'une mauvaise suite est-elle fausse.

   [Certaines idées de ce sujet sont inspirées de [Mad16].]

[Mad16] David Madore. Le lemme de Higman expliqué aux enfants, 2016. http://www.madore.org/ david/weblog/d.2016-04-26.2368.html#d .2016-04-26.2368.

[Wik18] Wikipedia. Higman's lemma, 2018. https://en.wikipedia.org/wiki/Higman's_lemma.