Mots finis, mots infinis & automates finis

Cours Master 2 RIF

Sandrine Julia

8 février 2019

Table des matières

Préambule

Ce cours propose une introduction à la combinatoire des mots classique, au sens de M. Lothaire ([10, 11]).
La combinatoire des mots a des origines qui se perdent un peu dans l'histoire. L'idée de considérer des suites de symboles est en fait très ancienne.
Généralement, on dit que la combinatoire des mots a débuté en 1906 avec le travail d'Axel Thue.
Dans les annéess 60 et 70, beaucoup de résultats classiques sur les mots étaient considérés comme faisant partie du *folklore* par les chercheurs travaillant en informatique théorique, et notamment dans l'école de M.P. Schützenberger en France.
Ce fut pour cette raison que le premier livre de Lothaire a paru [10]. En fait, le but de ce livre était principalement d'avoir un manuel de référence pour les résultats classiques sur les mots, ainsi que les principaux problèmes ouverts.
A partir de la parution du premier livre de Lothaire en 1983, on a pu constater une explosion de publications en combinatoire des mots. Rien que sur les mots *sturmiens*, on peut trouver des centaines de publications dans la littérature, et ceci est un phénomène qui ne connaît pas de répit.
Pourquoi donc ce succès de la combinatoire des mots ? Les raisons sont multiples. D'un côté, c'est un domaine d'interaction par excellence, car il se situe à la frontière entre les mathématiques discrètes et l'informatique théorique. Beaucoup de résultats combinatoires sur les mots ont permis de résoudre ou améliorer des problèmes techniques et industriels formalisés sous forme de séquences discrètes.
D'un autre côté, les problèmes que l'on trouve en combinatoire des mots présente un intérêt intellectuel intrinsèque.
Ce cours emprunte largement aux références bibliographiques [10, 11, 12], [4], [16].

Chapitre 1

Mots finis

1.1 Définitions de base

Un *alphabet* Σ est un ensemble fini non vide de symboles appelés *lettres*. La *taille* de l'alphabet notée $\mid \Sigma \mid$ est le nombre de ses éléments. Un alphabet binaire est un alphabet de taille 2. Un *mot* w sur Σ est une suite finie de lettres de Σ. La longueur d'un mot w notée $\mid w \mid$ est son nombre de lettres. Il existe un seul mot de longueur 0, c'est le *mot-vide* dénoté par ε. L'ensemble des mots de longueur n est noté Σ^n. Soit $w = \sigma_0\sigma_1\sigma_2...\sigma_{n-1}$ avec $n \geq 0$ un mot non vide, pour tout $i \in \{0, 1, 2, ..., n-1\}$, σ_i est la lettre à la *position* i dans w. Soient deux mots $w_1 = \sigma_0\sigma_1...\sigma_{n-1}$ et $w' = \sigma'_0\sigma'_1...\sigma'_{m-1}$, la *concaténation* de w et w' est le mot $ww' = \sigma_0\sigma_1...\sigma_{n-1}\sigma'_0\sigma'_1...\sigma'_{m-1}$. La concaténation étant une opération associative (mais pas commutative), l'ensemble des mots sur l'alphabet Σ est en fait le monoïde libre Σ^* engendré par Σ. L'élément neutre de Σ^* est bien sûr le mot vide. On note Σ^+ le semi-groupe égal à $\Sigma^* \setminus \{\varepsilon\}$. Un mot est synonyme de mot fini c'est-à-dire de longueur finie. Un *mot infini* est une séquence infinie de lettres. Un mot peut donc être infini à droite, infini à gauche ou bi-infini. On indique l'ensemble des mots infinis à droite, à gauche et bi-infinis avec respectivement les symboles Σ^ω, ${}^\omega\Sigma$ et ${}^\omega\Sigma^\omega$.

Remarque Un mot infini à droite est en fait une application de $\mathbb{N}$ dans Σ et un mot bi-infini une application de $\mathbb{Z}$ dans Σ. On trouve donc dans la littérature les symboles $\Sigma^{\mathbb{N}}$ et $\Sigma^{\mathbb{Z}}$ pour dénoter ces ensembles de mots.

Soit $w = uv$ avec $u \in \Sigma^*$ et $v \in \Sigma^*$, w est la concaténation des deux mots u et v. Le mot u est un *préfixe* de w tandis que v est un suffixe de w. u est un préfixe *propre* de w si u diffère de w. De la même façon, v est un suffixe *propre* de w si $v \neq w$. Un préfixe d'un suffixe de w est un *facteur* de w. Avec $w = \sigma_0\sigma_1...\sigma_{n-1}$, les facteurs sont les mots $\sigma_i...\sigma_j$ pour $0 \leq i \leq j \leq n-1$. Le facteur $\sigma_i...\sigma_j$ de w se note $w[i,j]$. On dénote respectivement par *Pref*(w), *Suff*(w) et *Fact*(w) l'ensemble des préfixes, des suffixes et des facteurs de w.

Lemme 1 *(Lévi, 1944) (extrait) (Propriété d'équidivisibilité)*
Pour tout mot x, y, z et t de Σ^*, si $xy = zt$ alors il existe un mot u de Σ^* tel que :

(1) soit $x = zu$ et $t = uy$

(2) soit $z = xu$ et $y = ut$

Exercice 1 Prouver *graphiquement* le lemme précédent.

Avec ces définitions, on peut également définir les suffixes d'un mot infini à droite ; il s'agira simplement des mots infinis à droite obtenus en effaçant un préfixe (fini) du mot.

Remarque Un mot (fini ou infini) binaire est un mot sur un alphabet binaire. Ceci signifie que dans un mot binaire, il y au *au plus* deux lettres distinctes. Ainsi, 000000 par exemple est considéré comme un mot binaire.

Théorème 1 *(König, 1926)*
Soit P un sous-ensemble infini de Σ^*. Alors il existe un mot infini w sur Σ tel que tout préfixe de w est un préfixe d'au moins un mot de P.

Démonstration Il existe au moins une lettre $w_1 \in \Sigma$ telle qu'un nombre infini de mots de P commencent par w_1. Il existe au moins une lettre $w_2 \in \Sigma$ telle qu'un nombre infini de mots de P commencent par w_1w_2, *etc.* On définit ainsi le mot $w_1w_2w_3...$ qui a la propriété voulue. $\square$

L'*image miroir* du mot $w = \sigma_0\sigma_1\sigma_2...\sigma_{n-1}$ est le mot $\tilde{w} = \sigma_{n-1}...\sigma_2\sigma_1\sigma_0$. A signaler que w^R dénote également l'image miroir (*reverse* en anglais).
Le *complément* d'un mot binaire est le mot $\overline{w}$ obtenu en échangeant les deux lettres entre elles dans w.

1.2 Bords et périodes

Un *bord* d'un mot w est un mot qui est à la fois préfixe propre et suffixe propre de w. Un bord u de w appartient donc à *Pref*$(w) \cap$ *Suff*(w) mais il diffère de w. Autrement dit, il existe deux mots v et v' dans Σ^+ tels que $w = vu = uv'$. Un mot est dit *sans bord* s'il n'a pas de bords non vides.

Exemple Les bords du mot $w = aabaabaa$ sont les éléments de $\{\varepsilon, a, aa, aabaa\}$.

Problème Déterminer une formule pour le nombre de mots sans bord est un problème ouvert.

Le bord d'un mot est le plus long de ses bords. La notion de *période* que nous présentons à présent est la notion duale de celle de *bord* comme l'atteste la propriété (iv) de la Proposition 1 suivante.

Un entier positif p vérifiant $0 < p \leq |w|$ est une *période* du mot w si $w_i = w_{i+p}$ pour tout $0 \leq i \leq \ |w| - p - 1$. Quand on parle de *la* période de w, on fait référence à la plus petite période de w, qu'on note souvent $\pi(w)$. Le préfixe de w de longueur $\pi(w)$ est appelé la *racine* de w et noté $\sqrt{w}$.

Exemple Les différentes périodes du mot $w = aabaa$ sont 3, 4 et 5. Sa période (la plus petite) est donc $\pi(w) = 3$. Les différentes périodes du mot $w = aabaabaa$ sont 3, 6, 7 et 8. Sa période est donc $\pi(w) = 3$.

Proposition 1 *Soit w un mot non vide de Σ^* et p un entier dans $\{1, ..., |w|\}$. Les propriétes suivantes sont équivalentes :*

(i) l'entier p est une période de w ;

(ii) il existe deux mots uniques $u \in \Sigma^*$ et $v \in \Sigma^+$ et un entier $k > 0$ tels que $w = (uv)^k u$ et $|uv| = p$;

(iii) il existe un mot t et un entier $k > 0$ tels que w est préfixe de t^k et $|\ t\ | = p$;

(iv) il existe trois mots u, v et x tel que $w = ux = xv$ et $|u| = |v| = p$;

(v) il existe un mot t tel que w est préfixe de tw et $|\ t\ | = p$.

Démonstration (partielle) (i) $\Rightarrow$ (ii) Soit k le dividende de la division euclidienne de $|w|$ par p. On obtient $k > 0$ car $w \neq \varepsilon$. Posons $u = w[0, r-1]$ et $v = w[r, p-1]$. On a bien $w = (uv)^k u$ avec $|uv| = p$. Dans tous les cas $v \in \Sigma^+$ et si w est une puissance alors $u = \varepsilon$. Supposons qu'il existe une autre possibilité (u', v', k') qui satisfait aux mêmes conditions que (u, v, k). On aurait $w = (u'v')^{k'} u'$ avec $|u'v'| = p$ et $v' \in \Sigma^+$. D'où $u'v' = uv$ et $|u'| < p$. Ainsi, $|wu'^{-1}| = k'p$ avec k' maximal. Donc $k' = k$ et il s'ensuit que $|u'| = |u|$ et donc que $u' = u$ et $v' = v$. Ainsi, cette décomposition est unique. $\square$

Exercice 2 Complétez la démonstration de la Proposition 1 du cas (ii) $\Rightarrow$ (iii) jusqu'à celui (v) $\Rightarrow$ (i) .

Remarque Un mot est de période 1 si et seulement si c'est la puissance d'une lettre $\sigma \in \Sigma$, c'est-à-dire s'il est de la forme $w = \sigma^n$, $n \geq 1$.

Lemme 2 *Un entier p est une période pour w si et seulement si w a un bord de longueur $\mid w \mid - p$.*

Démonstration De par l'équivalence des propriétés (i) et (iv) dans la Proposition 1. $\square$

Théorème 2 *(Fine et Wilf, 1965) (lemme de périodicité)*
Soit w un mot non vide ayant p et q comme périodes. Si

$$\mid w \mid \geq \; p + q - pgcd(p, q)$$

alors $pgcd(p, q)$ est une période de w.

Démonstration On procède par récurrence sur le $max\{p, q\}$. Le résultat est clairement vrai si $p = q = 1$ (et de façon générale quand $p = q$). Dorénavant, on suppose que $p > q$. D'après le (iv) de la Proposition 1, on a :

$$\begin{cases} w = ux \text{ avec } \mid u \mid = p \text{ et } x \text{ bord de } w \\ w = vz \text{ avec } \mid v \mid = q \text{ et } z \text{ bord de } w \end{cases}$$

On montre que $p - q$ est une période de z. En effet, on déduit de ce qui précède que x est un bord de z. Il s'ensuit que $|z| - |x|$ est une période de z. Or $|z| - |x| = (|w| - q) - (|w| - p) = p - q$ et donc $p - q$ est bien une période de z.
On montre que q est aussi une période de z. Puisque $p > q$ et $pgcd(p, q) \leq (p - q)$, on obtient que $q \leq p - pgcd(p, q)$. D'autre part, $p - pgcd(p, q) = p + q - pgcd(p, q) - q \leq |w| - q = |z|$. Il s'ensuit que $q \leq \; \mid z \mid$. Donc, d'après le (v) de la Prop. 1, la période q de w est aussi une période de z.
L'hypothèse du présent théorème est alors vérifiée pour les périodes $p - q$, q et le mot z. En effet, $(p - q) + q - pgcd(p - q, q) = p - pgcd(p - q, q) = p - pgcd(p, q) \leq |z|$.
On peut appliquer l'hypothèse de récurrence à $max\{p - q, p\}$ relativement au mot z et on obtient que le $pgcd(p, q)$ est une période de z.
Les conditions sur p, q et $pgcd(p, q) \leq (p - q)$ entraînent $q \leq |w|/2$. Comme par ailleurs $w = vz$ et que z est un bord de w, le mot v est un préfixe de z donc le mot z est un prefixe de vz. D'après le (v) de la Prop. 1, $|v|$ est période de z donc multiple de $pgcd(p, q)$.

Soit t le préfixe de w de longueur $pgcd(p, q)$. Alors, v est une puissance de t et z un préfixe d'une puissance de t. Le (iii) de la Prop. 1 assure que w est un préfixe d'une puissance de t. Donc $t = pgcd(p, q)$ est une période de w. $\square$

Exemple Le mot $w = aabaa$ n'a pas 1 comme période car $pgcd(3, 4) = 1$ certes mais $\mid w \mid = 5 < 3 + 4 - 1$.

La contraposée du théorème énonce que si un mot w a deux périodes p et q sans que $pgcd(p, q)$ soit une période, alors c'est que la longueur de w est strictement inférieure à $p + q - pgcd(p, q)$.

Exercice 3 Qu'en est-il des périodes du mot 01001010010 ?

1.3 Puissances, primitivité et conjugaison

Un *carré* est un mot de la forme $w = uu$ pour u un mot non vide. Un *cube* est un mot de la forme $w = uuu$ pour u un mot non vide. De façon générale, un mot est une *puissance-k^{eme}* s'il est de la forme $w = u^k$ pour un mot u non vide et un entier $k \geq 1$.

Pour un mot u de Σ^* et un entier k, la puissance k^{eme} de u notée u^k se définit inductivement par : $\begin{cases} u^0 = \varepsilon \\ u^k = u^{k-1}.u \end{cases}$

Un mot non vide est *primitif* s'il n'est puissance-k^{eme} que de lui-même. A noter que ε est exclus de l'ensemble des mots primitifs.

Problème Déterminer si le langage des mots primitifs sur Σ est *hors-contexte* est un problème ouvert.

Lemme 3 *Soient u et v deux mots. S'il existe deux entiers naturels non-nuls m et n tels que $u^m = v^n$ alors ils sont puissance d'un même mot z.*

Démonstration On se place dans le cas non trivial où ni u ni v n'est égal à ε. Si $min(m, n) = 1$, soit $z = v$ si $m = 1$, soit $z = u$ si $n = 1$. Si $min(m, n) > 1$, $|u|$ et $|v|$ sont des périodes du mot $w = u^m = v^n$. De plus, $|u| + |v| - pgcd(|u|, |v|) \leq |u| + |v| - 1 < |w|$. D'après le Théorème 2, u et v sont puissance de $z = w[0, pgcd(|u|, |v|) - 1]$, le préfixe de w de longueur $pgcd(|u|, |v|)$. $\square$

Lemme 4 *(Th. de Lyndon-Schützenberger)*
Si u et v sont deux mots non vides tels que $uv = vu$ alors ils sont puissance d'un même mot z.

Exercice 4 Démontrer le Lemme 4 par induction.

Lemme 5 *(lemme de primitivité)*
Un mot non vide est primitif si et seulement si il n'est un facteur de son carré qu'en tant que préfixe et suffixe.

Démonstration Si w est un mot non vide non primitif, il existe $u \in \Sigma^+$ et un entier $k \geq 2$ tels que $w = u^k$. Puisque w^2 se décompose sous la forme $u.u^k.u^{k-1}$, le mot w apparaît à la position $|u|$ dans w^2. Cela montre que tout mot non vide non primitif est un facteur de son carré sans en être seulement un préfixe ou un suffixe.
Réciproquement, soit w un mot non vide tel que son carré w^2 s'écrive sous la forme xwy avec $x, y \in \Sigma^+$. Pour une question de longueur, on a $|x| < |w|$. Ensuite, comme w est préfixe de xw, on obtient par la Proposition 1 que $|x|$ est une période de w. Ainsi, $|w|$ et $|x|$ sont des périodes de xw. D'après le Théorème 2, on en déduit que $p = pgcd(|w|, |x|)$ est également une période de xw. Maintenant, comme $p \leq |x| < |w|$, p est aussi une période de w. Et comme p divise $|w|$, on en déduit que w est de la forme t^k avec $|t| = p$ et $k \geq 2$. Cela montre que le mot w n'est pas primitif. $\square$

Exemple En appliquant le lemme de primitivité au mot $w = 011010$, on sait que w ne possède pas d'occurence « non triviale » (*i.e.* ni préfixe ni suffixe) dans son carré w^2. En revanche, le mot $v = 010101$ possède une occurence « non triviale » dans son carré v^2 car v n'est pas primitif : $v = (01)^3$.

Proposition 2 *Pour tout mot non vide, il existe un unique mot primitif dont il est puissance.*

Exercice 5 Prouver la proposition précédente.

Deux mots w et w' sont *conjugués* si il existe deux mots u et v tels que : $w = uv$ et $w' = vu$.

Proposition 3 *Deux mots non vides sont conjugués si et seulement si leurs racines le sont.*

Démonstration Soient w et w' deux mots non vides conjugués. On appelle x la racine de w et m son exposant et y la racine de w' et n son exposant. Ainsi, $w = x^m = uv$ et $w' = y^n = vu$. Il existe x' et $x'' \in \Sigma^+$ et $p, q \in \mathbb{N}$ tels que $x = x'x''$, $u = x^p x'$, $v = x'' x^q$ et $m = p + q + 1$. On en déduit que $w = x^p x' x'' x^q = (x'x'')^m$ et $w' = x'' x^q x^p x' = (x''x')^m$. Comme y est primitif, le Lemme 3 entraîne que $x''x'$ est une puissance de y, donc il existe un entier $k \in \mathbb{N}$ tel que $|x| = k|y|$. Par symétrie, il existe $l \in \mathbb{N}$ tel que $|y| = l|x|$. Il s'ensuit que $k = l = 1$, que $|x| = |y|$ et que $y = x''x'$. Les deux racines x et y sont bien conjuguées. La réciproque elle est immédiate. $\square$

Proposition 4 *(admis) Tout mot primitif a un conjugué sans bord.*

Proposition 5 *Deux mots non vides w et w' sont conjugués si et seulement si il existe un mot x tel que $wx = xw'$.*

Démonstration ($\Rightarrow$) Les mots u et v sont conjugués donc $w = uv$ et $w' = vu$. On prend $x = u$ et on obtient : $wx = (uv)u = uvu$ et $xw' = u(vu) = uvu$.
($\Leftarrow$) On a $wx = xw'$. On montre par récurrence sur $k \in \mathbb{N}$ que $w^k x = xw'^k$.

- vrai si $k = 0$: $w^0 x = \varepsilon x = x = x\varepsilon = x\omega'^0$ (vrai aussi si $k = 1$: $wx = xw'$) ;
- on suppose que pour tout $0 \leq i < k$, $w^i x = xw'^i$;
- $w^k x = ww^{k-1}x = wxw'^{k-1} = xw'w'^{k-1} = xw'^k$.

Soit $n \geq 1$ tel que $(n-1)|w| \leq |x| < n|w|$. Il existe deux mots u et $v \in \Sigma^*$ tels que $w = uv$, $x = w^{n-1}u$ et $vx = w'^n$ du fait que $w^n x = xw'^n$. On en déduit $w'^n = vw^{n-1}u = v(uv)^{n-1}u = (vu)^n$. Puisque $|w| = |w'|$, on a $w' = vu$, ce qui montre que w et w' sont conjugués. $\square$

1.3.1 Mots de Lyndon

Un *mot de Lyndon* est un mot primitif qui est minimal pour l'ordre lexicographique parmi tous ses conjugués. Autrement dit, w est un mot de Lyndon si et seulement si, pour toute factorisation de w en uv pour deux mots u et v non vides, on a $w <_{lex} v$. N'importe quel mot est le conjugué d'une puissance d'un mot de Lyndon.

Théorème 3 *(Chen, Fox, Lyndon, 1958) Tout mot non vide w admet une factorisation unique en une suite de mots de Lyndon décroissante pour l'ordre lexicographique :*

$$w = l_1^{k_1}\ l_2^{k_2}\ ....\ l_n^{k_n}$$

tels que $l_1 >_{lex} l_2 >_{lex} ... >_{lex} l_n$ et où, pour tout $1 \leq i \leq n$, l_i de Lyndon et $n_i \geq 1$.

Voici la suite des premiers mots de Lyndon triée par *ordre hiérarchique* (tri par longueur croissante et à longueur égale, tri par ordre lexicographique) :

0 1 01 001 011 0001 0011 0111 00001 00011 00101 00111 01011 01111

En 1983, un algorithme a été donné par J.-P. Duval pour décomposer un mot en mots de Lyndon.
En 1997, il a été montré que langage formé par les mots de Lyndon n'est pas hors-contexte. Sa clôture par permutation circulaire redonne le langage des mots primitifs.

1.4 Mots incomplétables

Pour un langage X inclus dans Σ^* donné, on dit qu'un mot w de Σ^* est *complétable* s'il est facteur d'un mot de X^*. X est *complet* si chaque mot $w \in \Sigma^*$ est complétable en un mot de X^*. On appelle U l'ensemble des *mots incomplétables*. Par définition :

$$U = \Sigma^* \setminus \mathit{Fact}(X^*)$$

Un mot w de U est *minimal incomplétable* si tous ses facteurs propres appartiennent à $\mathit{Fact}(X^*)$. On appelle M le sous-ensemble de U contenant les *mots minimaux incomplétables*.

Propriétés

(1) L'ensemble U de mots incomplétables par rapport à X est un idéal bilatère engendré par M :
$$U = \Sigma^* M \Sigma^*$$

(2) Soit w un mot de M, le préfixe $w\sigma^{-1}$ et le suffixe $\tau^{-1}w$, avec τ et σ les première et dernière lettres de w, appartiennent tout deux à $\mathit{Fact}(X^*)$.

Exemple Soit le langage $X = \{aaa, aaba, b, ba\}$. Le mot $aabaabaab \in U \setminus M$ tandis que $baabaab \in M$. En effet, $\mathit{Fact}(X^*)$ contient à la fois $baabaa$ et $aabaab$.

Parmi les mots incomplétables, on peut distinguer ceux de longueur minimale. On remarque que de tels éléments sont nécessairement dans M.

Exercice 6 : Prouver qu'un mot incomplétable de longueur minimale appartient à l'ensemble M.

Problème Soit X est un ensemble fini non complet. On appelle k la longueur maximale de ses mots. En toute généralité, on cherche l'ordre de grandeur d'un mot incomplétable de longueur minimale en fonction de k. En 1981, cette borne a été conjecturée en $2k^2$ [15]. Cette conjecture a été contredite depuis dans [**?**] sans que le fait qu'elle soit quadratique ne soit remis en question. Tout récemment dans [8], on doute du fait qu'elle soit seulement polynomiale.

1.5 Morphismes de mots

Soient Σ et A deux alphabets donnés. Un *morphisme* de Σ^* dans A^* est une application φ telle que :

$$\varphi(xy) = \varphi(x)\varphi(y)$$

$$\varphi(\varepsilon) = \varepsilon$$

Un morphisme sera donc entièrement défini par la donnée des images des lettres de Σ.

Exemple Soient $\Sigma = A = \{0, 1\}$, le morphisme est défini par $\varphi(0) = 10$ et $\varphi(1) = 01$. Ainsi, $\varphi^2(0) = 0110$ et $\varphi^3(0) = 10010110$ tandis que $\varphi^2(1) = 1001$, $\varphi^3(1) = 01101001$, ...

Un morphisme φ sur Σ est *effaçant* s'il existe une lettre σ tel que $\varphi(\sigma) = \varepsilon$. Un morphisme φ sur Σ est *prolongeable en* $\sigma \in \Sigma$ s'il existe un mot non vide $s \in \Sigma^*$ tel que : $\varphi(\sigma) = \sigma s$. Un mot fini (ou infini) w est un *point fixe* d'un morphisme φ si $\varphi(w) = w$.

Proposition 6 *Soit φ un endomorphisme sur Σ non effaçant, prolongeable en $\sigma \in \Sigma$ et soit s tel que $\varphi(\sigma) = \sigma s$.*

(1) pour tout entier $n \in \mathbb{N}$, $\varphi^{n+1}(\sigma) = \varphi^n(\sigma)\varphi^n(s)$;

(2) pour tout entier $n \in \mathbb{N}$, $\varphi^{n+1}(\sigma) = \sigma s \varphi(s)\varphi^2(s) \dots \varphi^n(s)$.

Exercice 7 Prouver par simple récurrence chacun des deux résultats de la Proposition 6 précédente.

Le problème PCP

Le Post Correspondance Problem est un problème célèbre qui s'énonce simplement sur les morphismes. Soient deux morphismes φ et ψ de Σ^* dans A^*. Existe-t-il un mot non vide x de Σ^* tel que $\varphi(x) = \psi(x)$?

Exercice 8 Considérons les deux morphismes φ et ψ de Σ^* dans A^*.

$$\begin{cases} \varphi(1) = 001 & \psi(1) = 00 \\ \varphi(2) = 11 & \psi(2) = 011 \\ \varphi(3) = 01 & \psi(3) = 000 \\ \varphi(4) = 010 & \psi(4) = 10 \end{cases}$$

Peut-on trouver un mot x de Σ^* tel que : $\varphi(x) = \psi(x)$?

Théorème 4 PCP *est un problème indécidable.*

L'idée pour prouver ce théorème consiste à effectuer une *réduction* à partir du *problème de la halte* d'une machine de Turing.

1.5.1 Mots de Fibonacci

Voici une famille célèbre de mots qui ont des propriétés intéressantes vis-à-vis des questions de périodicités et de répétitions.

Tout d'abord, les *nombres de Fibonacci* sont définis par récurrence :
Pour un mot u de Σ^* et un entier k, la puissance k^{eme} de u notée u^k se définit inductivement par : $\begin{cases} F_0 = 0 \\ F_1 = 1 \\ F_{n+2} = F_{n+1} + F_n & \text{pour } n \geq 0 \end{cases}$

Ces nombres vérifient des propriétés remarquables, en voici seulement quelques-unes.

Propriétés

(1) Pour tout entier naturel $n \geq 1$, $pgcd(F_{n+1}, F_n) = 1$;

(2) Pour tout entier naturel n, F_n est l'entier le plus proche de $\Phi^n/\sqrt{5}$ où $\Phi = (1+\sqrt{5})/2 = 1,61803...$ est le *nombre d'or* ;

(3) Pour tout entier naturel n, $F_n = \frac{1}{\sqrt{5}}(\Phi^n - \Psi^n)$ où Ψ dénote ici $(1-\sqrt{5})/2$.

Ensuite, les *mots de Fibonacci* sont définis par récurrence sur l'alphabet $\Sigma = \{0, 1\}$:

$$\begin{cases} f_0 = \varepsilon \\ f_1 = 1 \\ f_2 = 0 \\ f_{n+2} = f_{n+1} f_n \quad \text{pour } n > 0 \end{cases}$$

L'analogie entre nombres et mots de Fibonacci vient du remplacement de l'addition sur les entiers par la concaténation sur les mots. Ainsi, la suite des nombres de Fibonacci coïncide-t-elle avec celle des longueurs des mots de Fibonacci.

Exercice 9 Donner les huit premiers nombres et mots de Fibonacci et constater leurs liens.

Propriétés

(1) Pour tout entier $n \in \mathbb{N}$, $|f_n| = F_n$;

(2) Pour tout entier $n \geq 1$, $|f_n|_0 = F_{n-1}$;

(3) Pour tout entier $n \geq 2$, $|f_n|_1 = F_{n-2}$.

Exercice 10 Démontrer les trois propriétés précédentes.

Proposition 7 *(admis) Pour tout entier naturel $n \geq 1$, f_n est un mot primitif.*

Proposition 8 *(admis) La période de f_n est F_{n-1}. Le bord de $f_n = f_{n-2}$.*

Le *morphisme de Fibonacci* est défini comme suit :

$$\begin{cases} \varphi(0) = 01 \\ \varphi(1) = 0 \end{cases}$$

Exercice 11 Montrer que ce morphisme est injectif.

Propriété Pour tout entier naturel n, $f_{n+2} = \varphi^n(0)$.

1.6 Mots de Christoffel

Un mot de Christoffel est la discrétisation des segments de droite de pente rationnelle.

Si a et b sont premiers entre eux, le *chemin de Christoffel inférieur* de pente a/b est une suite de pas élémentaires horizontaux (0) ou verticaux (1) qui relient l'origine $(0,0)$ au point (b,a) vérifiant les conditions suivantes :

(i) c'est le chemin le plus près en dessous du segment de droite joignant ces deux points ;

(ii) il n'y a pas de points de $\mathbb{Z} \times \mathbb{Z}$ entre le chemin et le segment de droite.

On définit de façon analogue le *chemin de Christoffel supérieur*.

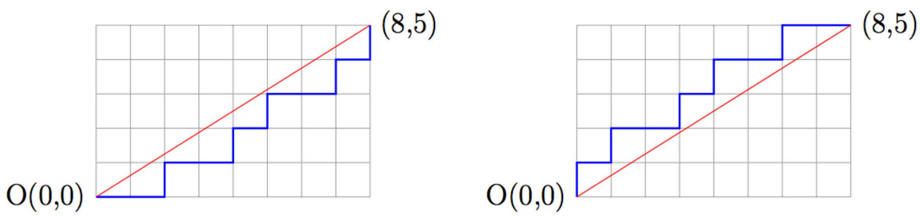

FIGURE 1.1 – Les chemins de Christoffel inférieur et supérieur de pente 5/8.

Pour un mot de Christoffel supérieur de pente a/b, le miroir de son complémentaire (obtenu en échangeant 0 et 1) est un mot de Christoffel inférieur de pente b/a et vice-versa.

Exercice 12 Trouver les mots de Christoffel de pente 3/5 et 5/3 et constater quels sont leurs liens.

Proposition 9 *Les mots de Christoffel sont des mots primitifs.*

Démonstration Si w est un mot de Christoffel de pente a/b, alors, $|w|_1 = a$ et $|w|_0 = b$ où $pgcd(a,b) = 1$. Cela entraîne que w ne peut pas être la puissance k^{eme} d'un autre mot avec $k \geq 2$. $\square$

Propriété Soit w un mot de Christoffel inférieur alors il existe un palindrome w' tel que le mot $w = 0w'1$.

Théorème 5 *Un mot w est un mot de Christoffel inférieur si et seulement si c'est un mot de Lyndon équilibré.*

Il s'ensuit que les chemins de Christoffel inférieur et supérieur sont conjugués.

Exemple Le mot de Christoffel inférieur de pente $2/3$ est égal à 00101. Par ordre lexicographique croissant, ses conjugués sont : 01001, 01010, 10010 et 10100. Le dernier étant le mot de Christoffel supérieur de même pente.

Chapitre 2

Mots infinis

On rappelle qu'un *mot infini* est une séquence infinie de lettres. Σ^ω, ${}^\omega\Sigma$ et ${}^\omega\Sigma^\omega$ sont respectivement les ensembles des mots infinis à droite, à gauche et bi-infinis. La notation $\Sigma^{\mathbb{N}}$ existe aussi pour les mots infinis à droite.

2.1 Définitions de base

Un mot infini w est *périodique* s'il existe un entier $p > 0$ tel que, pour tout entier $i \in \mathbb{N}$, $w_i = w_{i+p}$. L'entier p est une *période* de w.
Si un mot infini w est périodique de période p et si u est son préfixe de longueur p, alors w égale à u^ω, la concaténation à droite d'une infinité de mot u.
Un mot infini w est *ultimement périodique* s'il est de la forme uv^ω pour deux mots $u \in \Sigma^*$ et $v \in \Sigma^+$. Un mot infini qui n'est pas ultimement périodique est dit *apériodique*.
Un mot u est *facteur* d'un mot infini w s'il existe un mot (fini) p et un mot infini w' tel que $w = puw'$. Si $p = \varepsilon$, u est un préfixe de w. $\mathit{Fact}(w)$ désigne l'ensemble des facteurs du mot infini w et $\mathit{Fact}(w)_n$ l'ensemble de ses facteurs de longueur n. Un mot infini s est un suffixe d'un mot infini w s'il existe un mot fini p tel que $w = ps$. Le mot s est un *suffixe propre* de w si p diffère d'ε.
Un mot infini w est la limite d'une suite de mots $(u_n)_{n\geq 0}$ si pour tout préfixe p de w, il existe un entier n_p tel que, pour tout entier $n \geq n_p$, p est préfixe de u_n. On note :

$$w = \lim_{n\to\infty} u_n$$

En d'autres termes, tout préfixe de w est préfixe d'une infinité de u_n.
Un mot fini u est *récurrent* dans un mot infini w s'il apparaît infiniment souvent comme facteur de w. Un mot infini est *récurrent* si tous ses facteurs sont récurrents.

Un mot infini est *uniformément récurrent* si tous ses facteurs apparaissent infiniment souvent à « distance bornée ».
Un mot fini ou infini est *quasipériodique* si et seulement si il peut être couvert par des occurrences d'un mot q (autre que lui-même) alors appelé *quasi-période*. Par exemple, le mot $w = abaababaabaababaaba$ admet aba, $abaaba$ et $abaababaaba$ comme quasipériodes.

2.2 Mots morphiques

On rappelle qu'un mot infini (ou fini) w peut-être obtenu comme *point fixe* d'un morphisme φ si $\varphi(w) = w$. On peut à présent compléter la Proposition 6 précédente avec un troisième point relatif aux mots infinis :

Proposition 10 *Soit φ un endomorphisme sur Σ non effaçant, prolongeable en $\sigma \in \Sigma$ et soit s tel que $\varphi(\sigma) = \sigma s$.*

(1) pour tout entier $n \in \mathbb{N}$, $\varphi^{n+1}(\sigma) = \varphi^n(\sigma)\varphi^n(s)$;

(2) pour tout entier $n \in \mathbb{N}$, $\varphi^{n+1}(\sigma) = \sigma s\varphi(s)\varphi^2(s)\ldots\varphi^n(s)$;

(3) le mot infini $w = \sigma s\varphi(s)\varphi^2(s)\ldots\varphi^n(s)\ldots$ est égal à $\lim_{n\to\infty}\varphi^n(\sigma)$. De plus, il est l'unique mot débutant par σ point fixe de φ.

Un mot infini ainsi obtenu comme point fixe d'un morphisme se note $\varphi^\omega(\sigma)$, σ étant sa première lettre. On dit qu'il est *engendré* par un morphisme ou aussi que c'est un *mot (purement) morphique.*

Théorème 6 *Il est décidable si un mot morphique w est ultimement périodique. Il est décidable de savoir si un mot morphique est uniformément récurrent.*

2.2.1 Mot de Thue-Morse

On définit la suite $(t_n)_{n\geq 0}$ comme suit :

$$t_n = \begin{cases} 0 & \text{si } |bin(n)|_1 \text{ pair} \\ 1 & \text{si } |bin(n)|_1 \text{ impair} \end{cases}$$

Clairement, $t_0 = 0$, $t_1 = 1$, $t_2 = 1$, $t_3 = 0$, $t_4 = 1$, $t_5 = 0$, $t_6 = 0$, $t_7 = 1$ *etc.*
Le *mot de Thue-Morse* est un mot infini obtenu par concaténation infinie des mots de la suite $(t_n)_{n\geq 0}$,

$$\mathbf{t} = t_0t_1t_2t_3t_4t_5t_6t_7t_8\ldots = 011010011\ldots$$

Ce mot est un mot morphique. Il s'obtient comme point fixe du *morphisme de Thue-Morse* μ suivant :

$$\begin{cases} \mu(0) = 01 \\ \mu(1) = 10 \end{cases}$$

Ainsi, $\mu(0) = 01$, $\mu^2(0) = 0110$, $\mu^3(0) = 01101001$, $\mu^4(0) = 0110100110010110$ *etc*
Le mot de Thue-Morse correspond à $\mu^\omega(0)$.

Un *chevauchement* est un mot de la forme $w = \sigma u \sigma u \sigma$, $\sigma \in \Sigma$ et $u \in \Sigma$. C'est le fait que deux occurences d'un même mot $\sigma u \sigma$ se chevauchent. On dit qu'un mot est *sans chevauchement* s'il n'y a pas de chevauchement parmi ses facteurs.

A. Thue avait remarqué que tout mot de longueur ≥ 4 sur un alphabet binaire contenait forcément un carré. Il s'est demandé s'il existait un mot infini sans cubes et y a répondu positivement.

Théorème 7 *(Thue, 1906) Le mot de Thue-Morse* $\mathbf{t}$ *est un mot infini sans chevauchement. Par conséquent, il ne contient pas de cubes.*

Exercice 13 Les carrés sont-ils évitables pour un mot sur un alphabet à 3 lettres ? La réponse est oui, il reste à le prouver. *Indication : construire à partir du mot de Thue-Morse la suite qui compte le nombre de* 1 *entre deux* 0 *qui se suivent.*

2.2.2 Mot de Fibonacci

Le *mot de Fibonacci* est un mot infini obtenu comme point fixe du morphisme de Fibonacci déjà cité :

$$\begin{cases} \varphi(0) = 01 \\ \varphi(1) = 0 \end{cases}$$

$$\mathbf{f} = \varphi^\omega(0) = 0100101001001010010 10\ldots$$

Propriétés

(1) Le mot de Fibonacci n'est pas périodique ;

(2) Le mot de Fibonacci n'est pas ultimement périodique ;

(3) Le mot de Fibonacci est récurrent ;

(4) Le mot de Fibonacci est uniformément récurrent ;

(5) Le mot de Fibonacci est quasipériodique ;

(6) pour tout entier n, $\mathbf{f}$ a exactement $n+1$ facteurs différents de longueur n.

Un mot fini ou infini sur l'alphabet $\Sigma = \{0, 1\}$ est *équilibré* si pour tout couple de facteurs u et v de même longueur et pour toute lettre σ de Σ, on a :

$$|\ |u|_\sigma - |v|_\sigma\ | \leq 1$$

Lemme 6 *Le mot de Fibonacci* $\mathbf{f}$ *est équilibré.*

Un mot fini non vide ou infini est un *mot de Lyndon* si et seulement si il est strictement inférieur à tous ses suffixes propres non vides par l'ordre lexicographique.

Proposition 11 *On munit* $\Sigma = \{0, 1\}$ *de l'ordre* $0 <_{alph} 1$*. Le mot* $0\mathbf{f}$ *est un mot de Lyndon sur* Σ*.*

2.2.3 Mots sturmiens

Les mots sturmiens sont étudiés depuis au moins deux siècles mais ce terme a été introduit en 1944 par G.A. Hedlund. Ils ont des applications en mathématiques (algèbre, théorie des nombres), en physique (dynamique symbolique), en informatique (reconnaissance de motif, géométrie algorithmique). Les mots sturmiens admettent une multitude de définitions équivalentes : combinatoires, géométrique, *etc*

La *complexité factorielle* d'un mot infini w sur l'alphabet Σ est la fonction $f_w : \mathbb{N} \to \mathbb{N}$ définie ainsi, pour tout $n \geq 0$:

$$f_w(n) = |Fact(w) \cap \Sigma^n|$$

Voici une première définition combinatoire de ces mots : un mot infini est *sturmien* si c'est un mot apériodique et dont la complexité factorielle est minimale.

Exercice 14 Montrer que les mots ultimement périodiques sont caractérisés par une complexité factorielle bornée.

Théorème 8 *(Morse et Hedlund, 1940)*
Un mot infini est ultimement périodique si et seulement si il existe un entier n *tel que :*

$$f_w(n) \leq n$$

Si un mot w n'est pas ultimement périodique, alors, pour tout n, $f_w(n) \geq n+1$. Ainsi, les mots sturmiens sont les mots infinis apériodiques dont la complexité factorielle f_w vérifie :

$$f_w(n) = n + 1$$

Partant, les mots sturmiens sont définis sur un alphabet à deux lettres. Un mot sturmien, s'il n'est pas ultimement périodique, est uniformément récurrent, c'est-à-dire que n'importe lequel de ses facteurs apparaît infiniment souvent et que la distance entre deux occurrences consécutives est bornée.

Proposition 12 *Le mot de Fibonacci* $\mathbf{f}$ *est sturmien.*

Par conséquent, les mots sturmiens peuvent aussi être définis comme une généralisation du mot de Fibonacci. Soit $c_0, c_1, \ldots, c_n, \ldots$ une suite quelconque d'entiers tous strictement positifs sauf peut-être c_0. La limite $\mathbf{s}$ de la suite $(s_n)_{n\geq 0}$ suivante est un *mot sturmien standard* (la définition géométrique précisera plus tard quels mots sturmiens sont *standard*) :

$$\begin{cases} s_0 = b \\ s_1 = a \\ s_{n+1} = s_n^{c_{n-1}} s_{n-1} \quad \text{pour } n \geq 1 \end{cases}$$

Tous les mots sturmiens standard sont obtenus ainsi. Les mots finis d'une telle suite $(s_n)_{n\geq 0}$ sont appelés *mots finis standard* ou même *mots de Fibonacci généralisés.*

Les premiers mots à ne pas vérifier la condition du théorème de Fine et Wilf ont une longueur égale à $p+q-2$ dans le cas où les deux périodes p et q sont premières entre elles. De tels mots sont appelés *mots extrémaux de Fine et Wilf.*

Théorème 9 *(A. de Luca, F. Mignosi, 1994)*
L'ensemble des facteurs des mots extrémaux de Fine et Wilf coïncide avec l'ensemble des facteurs des mots sturmiens standard.

Les mots sturmiens admettent également une définition géométrique. Un mot sturmien peut être défini en considérant une suite de *sections* d'une grille faite par une demi-droite dont la pente est un nombre irrationnel. Une section horizontale est dénotée par la lettre b et une section verticale par la lettre a. Une section sur un angle le sera par ab ou ba. Les mots sturmiens représentés par une demi-droite partant de l'origine sont appelés *mots sturmiens standard.*

Théorème 10 *(Morse et Hedlund, 1940)*
Soit w un mot infini sur un alphabet binaire. Les conditions suivantes sont équivalentes :

(i) w est sturmien ;
(ii) w est mécanique de pente irrationnelle ;
(iii) w est équilibré et non ultimement périodique.

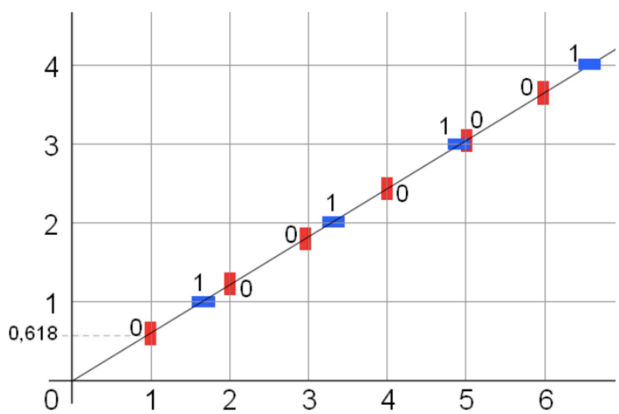

FIGURE 2.1 – Mot de Fibonacci obtenu par discrétisation de la droite de pente $\Phi - 1$.

La pente d'un mot w est la valeur, si elle existe :

$$\alpha = \lim_{n \to \infty} \frac{|w_1 w_2 w_3 ... w_n|_b}{n}$$

Si un mot infini est équilibré, sa pente est bien définie et elle appartient à $\mathbb{R}$. En outre :

Proposition 13 *Soit w un mot équilibré. La pente de w est rationnelle si et seulement si w est ultimement périodique.*

Corollaire 1 *Un mot infini équilibré est sturmien si et seulement si sa pente est irrationnelle.*

2.3 Mots lisses

En 1965, W. Kolakoski proposa le problème suivant : décrire une régle simple pour construire la suite :

$$\mathbf{K} = 12211212212211211221211212211211212212211212212...$$

Aujourd'hui, cette suite **K** est appelée le *mot de Kolakoski*. C'est en fait le seul mot sur l'alphabet $\Sigma = \{1, 2\}$ qui commence par 1 et qui coïncide avec son codage par plages (*Run-Length encoding*). En effet, en regroupant les lettres consécutives semblables avec un exposant, on obtient :

$$\mathbf{K} = 1^1 2^2 1^2 2^1 1^1 2^2 1^1 2^2 1^2 2^1 1^2 ...$$

Chapitre 3

Automates finis

Soit Σ un alphabet fini. Un *mot* u est une suite finie de lettres de Σ notée $u = u_1 ... u_n$, avec u_i dans Σ pour tout i compris entre 1 et n. Un ω-*mot* w est une suite infinie de lettres de Σ notée $w = w_1 w_2 w_3 ...$ avec w_i dans Σ, pour tout i strictement positif. On note ε le mot vide. Σ^* est l'ensemble des mots sur Σ et $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$. D'autre part, Σ^ω est l'ensemble des ω-mots sur Σ.
Les parties de Σ^* sont appelées des *langages*, celles de Σ^ω des ω-*langages*. Pour un langage L donné, nous présentons deux opérations :

$$L^* = \{\varepsilon\} \cup \{u_1 ... u_n,\ n > 0,\ \ \forall i,\ 1 \leq i \leq n,\ u_i \in L\}$$

Par convention, on pose : $L^0 = \{\varepsilon\}$. Clairement, $L^* = \bigcup_{n \geq 0} L^i$ et on note $L^+ = \bigcup_{n > 0} L^i$. De façon similaire, on définit à partir de L l'ω-langage suivant :

$$L^\omega = \{u_1 ... u_n ...,\ \forall i > 0,\ u_i \in L \setminus \{\varepsilon\}\}$$

Un ω-langage de la forme L^ω est dit ω-*puissance* de L. L est à la fois *générateur* de L^* et ω-*générateur* de L^ω.

Exemple Soit $\Sigma = a + b$ l'alphabet. $L = a^* b$ est un langage sur Σ. Son ω-puissance $L^\omega = (a^* b)^\omega$ est l'ω-langage qui contient tous les ω-mots ayant une infinité de fois la lettre b. Par abus de notation, un langage rationnel est noté comme étant égal à une expression régulière qui le décrit).

Soient L et L' deux langages sur Σ^*, ils vérifient les propriétés suivantes :

$$L^\omega = (L^*)^\omega = L^* L^\omega$$

$$(L + L')^\omega = (L + L')^* L^\omega + (L^* L')^\omega$$

3.1 Automates finis classiques

Dans cette partie, nous présentons les automates finis qui permettent de définir la famille des langages reconnaissables. Les ouvrages de référence sur la théorie des langages classiques sont [5, 7, 16], ceux sur la combinatoire des mots et des codes [1, 11, 2].
Un *automate fini* $\mathcal{A}$ est la donnée de $(\Sigma, Q, \delta, I, T)$ où Σ désigne un alphabet, Q un ensemble fini d'états, $I \subseteq Q$ un ensemble d'états initiaux, $T \subseteq Q$ un ensemble d'états terminaux et $\delta \subseteq (Q \times (\Sigma \cup \varepsilon) \times Q)$ un ensemble de transitions.
Pour tout $u \in \Sigma^*$, une *lecture* l de u dans $\mathcal{A}$ est une suite finie d'états $q_1, q_2, ..., q_n$ telle que $\forall i,\ 1 \leq i < n,\ (q_i, u_i, q_{i+1}) \in \delta$.
Un mot $u \in \Sigma^*$ est *reconnu* par $\mathcal{A}$ s'il existe une lecture *réussie* l de u, c'est-à-dire telle que $q_1 \in I$ et $q_n \in T$. Le *langage reconnu* par $\mathcal{A}$ est l'ensemble des mots admettant une lecture réussie dans $\mathcal{A}$. On le note $L(\mathcal{A})$.

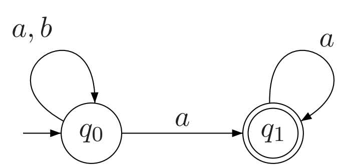

FIGURE 3.1 – L'automate fini $\mathcal{A}$

Exemple Soit $\mathcal{A}$ l'automate $(\Sigma, Q, \delta, I, T)$ tel que $Q = \{q_0, q_1\}$, $I = \{q_0\}$, $T = \{q_1\}$ et $\delta = \{(q_0, a, q_0), (q_0, b, q_0), (q_0, a, q_1), (q_1, a, q_1)\}$. Figure 3.5, le langage reconnu par $\mathcal{A}$ est $L(\mathcal{A}) = (a + b)^*a$.

3.1.1 Reconnaissabilité et rationalité

Les *langages rationnels* de Σ^* forment l'ensemble $Rat(\Sigma^*)$ défini comme étant la plus petite famille de parties de Σ^* telle que :

- $\varnothing \in Rat(\Sigma^*)$
- $\forall \sigma \in \Sigma \quad \{\sigma\} \in Rat(\Sigma^*)$
- $\forall A, B \in Rat(\Sigma^*)$:
  $A \cup B \in Rat(\Sigma^*)$
  $AB \in Rat(\Sigma^*)$
  $A^* \in Rat(\Sigma^*)$.

L'ensemble des *langages rationnels* est aussi la clôture par union, produit et * de l'ensemble des langages finis. Ces langages peuvent être décrits par une *expression* *rationnelle.*

Théorème 11 *(Kleene) [9]*
Soit Σ est un alphabet fini, un ensemble inclus dans Σ^* est reconnaissable par un automate fini si et seulement si il est rationnel.

3.1.2 Déterminisme

Un automate fini $\mathcal{A}$ est *déterministe* si son ensemble I d'états initiaux est singleton, s'il ne contient pas d'ε-transitions et si pour tout $q \in Q$ et pour tout $\sigma \in \Sigma$, il existe au plus un triplet $(q, \sigma, q') \in \delta$.
Un langage reconnu par un automate fini est reconnaissable par un automate fini déterministe et on a l'égalité entre $Rec(\Sigma^*)$ et $Drec(\Sigma^*)$.

Proposition 14 *[5] Soit L un langage de Σ^*. Les propositions suivantes sont équivalentes :*
(i) $L \in Rec(\Sigma^*)$
(ii) il existe un automate fini $\mathcal{A}$ tel que $L = L(\mathcal{A})$.
(iii) il existe un automate fini déterministe $\mathcal{A}$ tel que $L = L(\mathcal{A})$.

Exemple L'automate fini $\mathcal{A}$ précédent n'est pas déterministe. En revanche, sur la Figure 3.2, l'automate fini $\mathcal{B} = (\Sigma, Q, \delta, I, T)$ tel que $Q = \{q_0, q_1\}$, $I = \{q_0\}$, $T = \{q_1\}$ et $\delta = \{(q_0, a, q_1), (q_0, b, q_0), (q_1, a, q_1), (q_1, b, q_0)\}$ est déterministe. $\mathcal{B}$ est la version déterministe de $\mathcal{A}$ dans la mesure où $L(\mathcal{B}) = L(\mathcal{A}) = (a + b)^*a$.

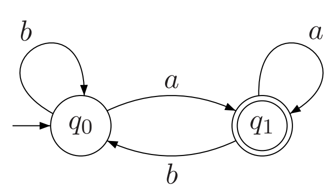

FIGURE 3.2 – L'automate fini déterministe $\mathcal{B}$

3.1.3 Automates finis et logique

Les langages rationnels de mots infinis sont nés de préoccupations logiques. En 1960, Büchi a proposé une *méthode effective* pour décider de la logique monadique du 2^{nd} ordre [3]. Il présentait alors cette méthode comme *typique de la théorie des automates.*

La logique monadique *faible* du 2^{nd} ordre avec un successeur correspond aux automates finis finitaires alors que la logique monadique du 2^{nd} ordre avec un successeur correspond aux automates de Büchi. Tout ceci constitue les bases théoriques du **model-checking**.
Un langage peut être défini par une propriété et le problème est de savoir quelles sont les propriétés qui définissent une classe de langages, par exemple ici celle des langages rationnels de mots finis.
On se place sur l'alphabet $\Sigma = \{a, b\}$. Voici quelques propriétés qui caractérisent les mots qui forment des langages particuliers :

(1) toute occurrence de a est suivie (de près ou de loin) par un b

(2) il y a exactement un a.

(3) la première lettre est un a

(4) toute occurrence de a est immédiatement suivie d'un b.

(5) le nombre de lettres est pair

A noter que la propriété *le nombre de a est égal à celui de b* ne définit pas un langage rationnel.

Un langage naturel pour décrire de telles propriétés comme celles ci-dessus est la **logique du 1$^{\mathrm{er}}$ ordre sur les mots**. Les *variables* x, y, z... font référence au rang ou position d'une lettre dans un mot. Les prédicats comme $a(x)$ sont vrais dès qu'il y a une lettre a au rang x d'un mot. La formule $y > x$ est vraie si le rang y est quelque part à droite du rang x. La première des propriétés précédentes correspond à la formule :

$$\forall x \;\; (\; a(x) \;\; \Rightarrow \;\; \exists y \; ((y > x) \wedge b(x)) \;)$$

La deuxième propriété peut être exprimée ainsi :

$$\forall x \; \forall y \;\; (a(x) \wedge a(y)) \;\; \Rightarrow \;\; x = y$$

Exercice 15 Appelons $\mathit{First}(x) \triangleq \forall y \; (x = y) \vee (x < y)$. Cette formule s'évalue à *Vrai* sur un x donné si et seulement si x est le premier rang. Exprimez la formule correspondant à la 3^{eme} propriété ci-dessus.

Exercice 16 Donnez une formule correspondant à la 4^{eme} propriété ci-dessus en commençant par exprimer une formule pour vérifier que y est bien la position qui suit immédiatement x.

Bien que décrivant un langage rationnel de mots finis, la 5^{eme} propriété n'est pas exprimable dans cette logique.

Pour l'instant, les variables dont nous disposons ne portent que sur les positions des lettres dans un mot. En **logique monadique du $\mathbf{2^{nd}}$ ordre**, nous nous autorisons des variables X, Y, Z... portant sur des ensembles de positions. On peut quantifier ces variables et utiliser l'appartenance.
A présent, on peut reformuler la 5^{eme} propriété dès lors que l'ensemble des positions est divisé en 2 ensembles disjoints X et Y :

- la première position est dans X
- la seconde est dans Y
- pour chaque position dans X, la suivante, si elle existe, est dans Y
- pour chaque position dans Y, la suivante, si elle existe, est dans X

Voici la formule de logique monadique de 2^{nd} ordre qui exprime cette propriété :

$$\begin{aligned}\exists X\ \exists Y\ \ &(\ \forall x\ (x \in X)\ \vee\ (x \in Y)\\ &\wedge\ \forall x\ \ (x \in X)\ \iff\ (x \notin Y)\\ &\wedge\ \forall x\ \mathit{First}(x) \Rightarrow (x \in X)\\ &\wedge\ \forall x\ \mathit{Last}(x) \Rightarrow (x \in Y)\\ &\wedge\ \forall x\ \forall y\ (\ (x \in X) \wedge (y = x+1)\ ) \Rightarrow (y \in Y)\\ &\wedge\ \forall x\ \forall y\ (\ (x \in Y) \wedge (y = x+1)\ ) \Rightarrow (y \in X)\ \ )\end{aligned}$$

Cette logique étendue qui inclut l'usage de variables d'ensemble et l'appartenance est la logique monadique de 2^{nd} ordre sur les mots. On la note MSO ou aussi S1S pour un successeur. La classe des langages définissables par les formules de la logique monadique du 2^{nd} ordre sont précisément les langages rationnels.

Théorème 12 *(Büchi [3],Elgot [6]) Un langage L est rationnel si et seulement si il peut être décrit par une formule de MSO.*

Comment passer d'un automate fini à une formule de MSO ? Considérons d'abord le cas particulier de l'automate de la Figure 3.3.

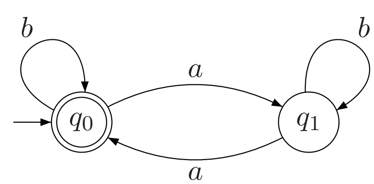

FIGURE 3.3 – Un automate fini $\mathcal{A}$

La lecture du mot *ababaa* se décompose comme suit :

$$q_0 \xrightarrow{a} q_1 \xrightarrow{b} q_1 \xrightarrow{a} q_0 \xrightarrow{b} q_0 \xrightarrow{a} q_1 \xrightarrow{a} q_0$$

La lecture d'un mot de n lettres parcourt $n+1$ états. En faisant abstration pour l'instant du dernier état visité, on tente de décrire par une formule à la fois la lettre lue et l'état qui précède cette lecture. L'idée consiste à utiliser des variables X_q du 2^{nd} ordre pour décrire chacun des états q. Notre exemple requiert seulement les variables d'ensemble X_{q_0} et X_{q_1}.

(1) La première position appartient à X_{q_0} car q_0 est l'état initial.

(2) Si x et $x+1$ sont deux positions consécutives dans un mot donné, et que la lettre en position x est un a, on obtient que x et $x+1$ appartiennent respectivement aux ensembles X_p et X_q si $\delta(p,a)=q$.

(3) Si x est la dernière position d'un mot w, que la lettre à cette position est σ et que x appartient à X_q alors $\delta(q,\sigma)$ doit appartenir à F.

Voici la formule de la logique MSO qui décrit l'automate $\mathcal{A}$:

$$\begin{array}{ll}
\exists X_{q_0}\ \exists X_{q_1} & (\ \forall x\ (x \in X_{q_0})\ \vee\ (x \in X_{q_1}) \\
& \wedge\ \forall x\ (x \in X_{q_0})\ \Longleftrightarrow\ (x \notin X_{q_1}) \\
& \wedge\ \forall x\ \mathit{First}(x) \Rightarrow (x \in X_{q_0}) \\
& \wedge\ \forall x\ \forall y\ ((x \in X_{q_0}) \wedge a(x) \wedge (y = x+1))\ \Rightarrow (y \in X_{q_1}) \\
& \wedge\ \forall x\ \forall y\ ((x \in X_{q_0}) \wedge b(x) \wedge (y = x+1))\ \Rightarrow (y \in X_{q_0}) \\
& \wedge\ \forall x\ \forall y\ ((x \in X_{q_1}) \wedge a(x) \wedge (y = x+1))\ \Rightarrow (y \in X_{q_0}) \\
& \wedge\ \forall x\ \forall y\ ((x \in X_{q_1}) \wedge b(x) \wedge (y = x+1))\ \Rightarrow (y \in X_{q_1}) \\
& \wedge\ \forall x\ (\mathit{Last}(x) \Rightarrow\ ((x \in X_{q_0} \wedge b(x)) \vee (x \in X_{q_1} \wedge a(x))))\ )
\end{array}$$

Les deux premières lignes indiquent que chaque position est assortie d'un des 2 états q_0 ou q_1. La troisième ligne fait en sorte que la première position soit bien assortie de l'état initial. Les quatre lignes suivantes encodent les quatre transitions de l'automate. La dernière ligne assure que l'état allant avec la dernière position mène à un état final de F sur la dernière lettre.

En toute généralité, pour l'automate fini $\mathcal{A} = (\Sigma, \{q_0, ..., q_k\}, \delta, q_0, F)$, les modèles (finitaires) de la formule générale ci-dessous sont acceptés par $\mathcal{A}$:

$$\begin{array}{ll}
\exists X_{q_0}\ ...\ \exists X_{q_k} & (\ \forall x\ \bigvee_{0 \le i \le k}\ x \in X_{q_i} \\
& \wedge\ \forall x\ \bigwedge_{i \neq j} (x \in X_{q_i})\ \Leftrightarrow\ \neg(x \in X_{q_j}) \\
& \wedge\ \forall x\ \mathit{First}(x) \Rightarrow (x \in X_{q_0})
\end{array}$$

$$\begin{array}{l} \bigwedge_{\delta(p,a)=q} \ \forall x\ \forall y((x \in X_p) \wedge a(x) \wedge (y = x+1)) \ \Rightarrow (y \in X_q) \\ \wedge\ \ \forall x\ Last(x) \Rightarrow\ (\bigvee_{\delta(p,a)\in F} ((x \in X_p \wedge a(x))) \end{array}$$

Les deux premières lignes partionnent l'ensemble des positions en fonction de l'état qui va avec. Le reste a déjà été décrit sur l'exemple. En conclusion, un mot satisfait cette formule si et seulement si il est accepté par l'automate fini $\mathcal{A}$.

Exercice 17 Trouvez l'automate fini minimal reconnaissant le langage des mots binaires contenant le motif 00. Traduisez-le en une formule de la logique MSO.

Ceci est une simple évocation du premier sens de la démonstration du Théorème 12. Resterait à évoquer le passage réciproque d'une formule MSO à un automate fini (cf. [18, 20]). Les langages rationnels de mots infinis étudiés dans ce cours ne sont rien d'autres que les modèles infinitaires des formules de la logique MSO pour décrire un automate.

3.2 Langages rationnels de mots infinis

Nous présentons la théorie des langages rationnels de mots infinis et des automates finis qui les reconnaissent : les *automates de Büchi* [19, 17, 14].

3.2.1 ω-mots et ω-langages

Soit Σ un alphabet fini. Un mot infini ou ω-*mot* w est une suite infinie de lettres de Σ notée $w = w_1w_2w_3$ ... . Pour tout $i > 0$, w_i est la i^{eme} lettre de w. Σ^ω est l'ensemble des ω-mots sur l'alphabet Σ. On note $\Sigma^\infty = \Sigma^* \cup \Sigma^\omega$ le monoïde qui regroupe les mots finis et infinis sur Σ.

Soit L et L' deux éléments de Σ^∞, on définit le *produit de concaténation*

$$L.L' = \{uv/\ u \in L, v \in L'\}$$

L'ensemble des parties de Σ^∞ dénoté $2^{\Sigma^\infty} = \mathcal{P}(\Sigma^\infty)$ est un monoïde dont $\{\varepsilon\}$ est l'élément neutre. Soit $L \subseteq \Sigma^\infty$, on note $L_{fin} = L \cap \Sigma^*$ et $L_{inf} = L \cap \Sigma^\omega$. Clairement, $L = L_{fin} \cup L_{inf}$.
On en déduit :

$$L.L' = L_{fin}.L'_{fin}\ \cup\ L_{fin}.L'_{inf}\ \cup\ L_{inf}$$

$$L^* = L^*_{fin}\ \cup\ L^*_{fin}\ L_{inf}$$

$$(L_{fin})^* = (L^*)_{fin}$$

Les parties de Σ^ω sont appelées des ω-*langages*. Par analogie avec l'étoile de Kleene, on définit l'opérateur de concaténation infinie $^\omega$. A tout langage $L \subseteq \Sigma^\infty$, la puissance $^\omega$ associe l' ω-langage de Σ^ω :

$$L^\omega = \{u_1...u_n..., \ \forall i > 0, \ u_i \in L \setminus \{\varepsilon\}\}$$

Par abus de notation, un langage rationnel est noté comme étant égal à une expression régulière qui le décrit. On s'autorise d'ores et déjà l'opération $^\omega$ pour décrire les ω-langages rationnels présentés plus loin.

Exemple Soit Σ l'alphabet $\{a, b\}$. $L = a^*b$ est un langage sur Σ. Son ω-puissance $L^\omega = (a^*b)^\omega$ est l'ω-langage qui contient tous les ω-mots ayant une infinité de fois la lettre b.

Exercice 18 Sur l'alphabet $\{a, b\}$, décrivez le langage des mots infinis ne contenant pas la lettre a isolée (entre 2 b ou en début de mot).

Pour tout $w \in \Sigma^\omega$, le mot $w \triangleq u_1...u_n...$ si et seulement si $\{u_1...u_n\}_{n>0} \subseteq Pref(w)$.

Soient L et M deux langages sur Σ^∞ avec $L = L_{fin} \cup L_{inf}$. Ils vérifient les propriétés suivantes :

$$L^\omega = (L_{fin})^* L_{inf} \cup (L_{fin})^\omega$$

$$L^\omega = L^* L^\omega = L^+ L^\omega = L^i L^\omega = L^\omega L^\omega \text{ pour } i > 0$$

$$L^\omega = (L^*)^\omega = (L^+)^\omega = (L^i)^\omega = (L^\omega)^\omega \text{ pour } i > 0$$

$$(L^\omega)^* = \{\varepsilon\} \cup L^\omega$$

$$(L \cup M)^\omega = (L \cup M)^* L^\omega \cup (L^* M)^\omega$$

$$(L.M)^\omega = L(ML)^\omega$$

$$\text{si } M \neq \varnothing, \ L^\omega M = L^\omega$$

Exercice 19 Prouvez quelques-unes des propriétés précédentes, par exemple que :

$$(L \cup M)^\omega = (L \cup M)^* L^\omega \cup (L^* M)^\omega \quad .$$

Exemple $(a + b)^\omega = (a + b)^* b^\omega + (b^* a)^\omega$ où le nombre de a est fini puis infini. De plus, $(b^* a)^\omega = a^*(b^+ a^+)^\omega + (b^* a)^* a^\omega$ avec cette fois le nombre de b infini puis fini.

3.2.2 Rationalité

On s'apprête à définir la famille des parties rationnelles de mots finis ou infinis. Pour cela, notons $Rat(\Sigma^*)$ la classe des langages rationnels et $Fin(\Sigma^*)$ son sous-ensemble contenant les parties de cardinalité finie de Σ^*.

Définition $Rat(\Sigma^\infty)$ est la plus petite famille de langages de Σ^∞ contenant $Fin(\Sigma^*)$ et fermée par union, concaténation, étoile et puissance $^\omega$.

Proposition 15 *[5] Les assertions suivantes sont équivalentes :*
(i) $L \in Rat(\Sigma^\infty)$
(ii) $L = \bigcup_{i=1}^n (A_i + B_i C_i^\omega)$ *avec pour tout* $1 \leq i \leq n$, $A_i, B_i, C_i \in Rat(\Sigma^*)$.

Démonstration (ii) $\Rightarrow$ (i) $Rat(\Sigma^*) \subseteq Rat(\Sigma^\infty)$ donc pour tout i, $A_i, B_i, C_i \in Rat(\Sigma^\infty)$ donc L aussi.
(i)$\Rightarrow$ (ii) Par induction. Clairement, tout élément de $Fin(\Sigma^*)$ peut être mis sous la forme de (ii). On montre alors que le type (ii) est stable pour +, ., * et $^\omega$.
+ : clair
. : $L.M = \bigcup_{i=1}^n \bigcup_{j=1}^p (L_i M_j)$. Il suffit de montrer que $L_i M_j$ est de type (ii) comme suit : $L_i M_j = A_i A'_j + B_i C_i^\omega + (A_i B'_j) C_j'^\omega$.
* : $L = \bigcup_{i=1}^n A_i + B_i C_i^\omega = A + \bigcup_{i=1}^n B_i C_i^\omega$ avec $A \in Rat(\Sigma^*)$. Donc $L^* = A^* + \bigcup_{i=1}^n A^* B_i C_i^\omega$ qui est bien du type (ii).
$^\omega$: $L^\omega = A^\omega + \bigcup_{i=1}^n A^* B_i C_i^\omega$ est bien du type (ii). □

Définition [5]

$$Rat(\Sigma^\omega) \triangleq_1 \{L_{inf} / L \in Rat(\Sigma^\infty)\}$$

$$Rat(\Sigma^\omega) \triangleq_2 \{L \subseteq \Sigma^\omega / L \in Rat(\Sigma^\infty)\}$$

$$Rat(\Sigma^\omega) \triangleq_3 \{L = \bigcup_{i=1}^n A_i B_i^\omega / \ A_i, B_i \in Rat(\Sigma^*)\}$$

Exercice 20 Vérifiez que les 3 définitions précédentes de $Rat(\Sigma^\omega)$ sont bien équivalentes.

La clôture $Rat(\Sigma^\infty)$ de la classe des langages finis par : union, produit, * et $^\omega$ est constituée de parties dans $\Sigma^\infty = \Sigma^* \cup \Sigma^\omega$. Parmi elles, les sous-ensembles de Σ^ω forment donc la classe des ω-langages rationnels notée $Rat(\Sigma^\omega)$. Ceci justifie d'étendre le formalisme des expressions régulières à la puissance $^\omega$ pour décrire les

ω-langages rationnels.

Un mot w est *ultimement périodique* s'il est de la forme $\alpha\beta^\omega$ avec $\alpha \in \Sigma^*$ et $\beta \in \Sigma^+$.

Proposition 16 *Soit $w \in \Sigma^\omega$. L'ω-langage-singleton $\{w\}$ est dans $Rat(\Sigma^\omega)$ si et seulement si le mot w est ultimement périodique.*

Démonstration ($\Leftarrow$) clair.
($\Rightarrow$) $\{w\} = \bigcup_{i=1}^{n} A_i B_i^\omega$. Soit $\alpha \in A_1$ et $\beta \in B_1$, alors $\alpha\beta^\omega \in \{w\}$. D'où $w = \alpha\beta^\omega$. $\square$

Exercice 21 Donnez un exemple de mot w non-ultimement périodique pour lequel le singleton $\{w\}$ n'est donc pas dans $Rat(\Sigma^\omega)$.

Propriétés

(1) Si w et $w' \in \Sigma^\omega$ alors $(w = w') \Leftrightarrow (Pref(w) = Pref(w'))$
(2) $Pref(L^*) = Pref(L^\omega)$
(3) Si L, $L' \in \Sigma^\omega$ alors $(L = L') \Rightarrow (Pref(L) = Pref(L'))$ - réciproque fausse -
(4) Si $L \in Rat(\Sigma^\omega)$ alors $Pref(L) \in Rat(\Sigma^*)$ - réciproque fausse -.

Exercice 22 Donnez des contre-exemples qui prouvent que les réciproques de (3) et de (4) sont bien fausses.

3.2.3 Limite

Outre la puissance $^\omega$, la limite et l'adhérence (non traitée ici) constituent deux autres moyens d'associer un ω-langage à un langage donné.
Soit $u_1 < u_2 < u_3$ ... une suite strictement croissante (pour l'ordre préfixe) d'éléments de Σ^*, elle admet une borne supérieure $w \in \Sigma^\omega$ telle que $\forall n \geq 0, \exists i \geq 0, \quad w[n] = u_i[n]$.
Soit L un langage, la *limite* de L (cf. Figure 3.4) se note $\overrightarrow{L}$ et est l'ensemble des bornes supérieures des suites strictement croissantes contenues dans L.

Définition

$$\overrightarrow{L} = \{w \in \Sigma^\omega, \ Card(Pref(w) \cap L) est\ infini\ \}$$

La limite d'un langage L n'est autre que l'ensemble des ω-mots qui ont une infinité de préfixes dans L. La limite d'un langage fini est vide, de même que celle d'un langage *préfixe* (*prefix-free* en anglais).

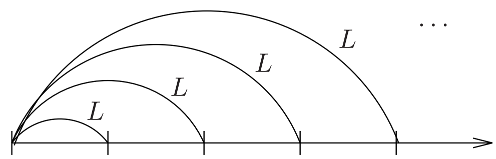

FIGURE 3.4 – Un ω-mot de la limite de L

Exemple Si $L = a^*b$, $\overrightarrow{L} = \varnothing$. Si $L = (ab)^+$ alors $\overrightarrow{L} = (ab)^\omega$. Si $L = (a^*b)^+ = (a+b)^*b$, L regroupe les mot se terminant par b. Sa limite $\overrightarrow{L} = (a^*b)^\omega$ est l'ensemble des ω-mots contenant une infinité de b.

Exercice 23 Trouvez les limites $\overrightarrow{L}$ des langages successifs : $L = ba^*$, $L = (ba^*)^*$, $L = b^*a$, $L = (ab)^*a$, $L = a^*b^+$, $L = a^*b^*$, $L = (ab+c)^*a$.

Tous les ω-langages ne sont pas des limites, ni même tous les ω-langages rationnels. Nous étudierons la classe des ω-langages qui sont des limites à la Section 3.3.2.

Exemple L'ω-langage $L = (a+b)^*a^\omega$ n'est pas une limite.

Remarques
— si $w \in \Sigma^\omega$ alors $w = \overrightarrow{Pref(w)}$
— si $L \subseteq \Sigma^\omega$ alors $L \subseteq \overrightarrow{Pref(L)}$
— par conséquent, $\Sigma^\omega = \overrightarrow{\Sigma^*}$

La limite d'un langage rationnel est un ω-langage rationnel car finalement, ils sont reconnus par *le même automate* comme on le verra ultérieurement.

Propriétés

(1) $\overrightarrow{L \cup M} = \overrightarrow{L} \cup \overrightarrow{M}$

(2) $L.\overrightarrow{M} \subseteq \overrightarrow{L.M}$

(3) $\overrightarrow{L.M} \subseteq \overrightarrow{L} \cup L.\overrightarrow{M}$

(4) $L^\omega \subseteq \overrightarrow{L^*}$ - réciproque fausse -

Exemple Avec $L = a^*$ et $M = b^*$, $\overrightarrow{L \cup M} = \overrightarrow{L} \cup \overrightarrow{M} = a^\omega + b^\omega$. $L.\overrightarrow{M} = a^*b^\omega \subsetneq \overrightarrow{L.M} = a^\omega + a^*b^\omega$. Ici, $\overrightarrow{L.M} = \overrightarrow{L} \cup L.\overrightarrow{M}$

Exercice 24 Adaptez les propriétes précédentes au cas où L est un langage préfixe.

Proposition 17 *Soit L un langage. Il existe la relation suivante entre puissance $^{\omega}$ et limite :*

$$\overrightarrow{L^*} = L^\omega + L^* \overrightarrow{L}$$

On commence par présenter le *lemme d'itération infinie* qui est nécessaire à la démonstration de la Proposition 17.

Lemme 7 *Pour tout $L \subseteq \Sigma^+$ et tout $K \subseteq \Sigma^\omega$, si $K \subseteq LK$ alors $K \subseteq L^\omega$.*

Exercice 25 Démontrez le *lemme d'itération infinie*.

Démonstration (Prop. 17) ($\supseteq$) D'une part, on sait que $L^\omega \subseteq \overrightarrow{L^*}$. D'autre part, $L^* \overrightarrow{L} \subseteq \overrightarrow{L^* L} = \overrightarrow{L^+} = \overrightarrow{L^*}$.
($\subseteq$) On pose $K = \overrightarrow{L^*} \setminus L^* \overrightarrow{L}$. A-t-on $K \subseteq L^\omega$? L'idée est de commencer par prouver que $K \subseteq LK$ et d'appliquer le Lemme 7. On a $\overrightarrow{L^*} = \overrightarrow{LL^*} = \overrightarrow{L} + L.\overrightarrow{L^*}$ or si $w \in K$, $w \notin \overrightarrow{L}$ donc $w \in L.\overrightarrow{L^*}$. On peut supposer que $\varepsilon \notin L$. L'ω-mot $w = uv$ avec $u \in L$ et $v \in \overrightarrow{L^*}$. On montre alors que $v \in K$, il faut donc montrer que $v \notin L^* \overrightarrow{L}$. Par l'absurde, si $v \in L^* \overrightarrow{L}$ alors $w \in L^+ \overrightarrow{L} \subseteq L^* \overrightarrow{L}$: impossible. Donc $v \in K$ et $w \in LK$ soit $K \subseteq LK$. D'après le lemme, $w \in L^\omega$. $\square$

3.3 Automates de Büchi

Dans cette partie, nous élargissons la reconnaissance par automates aux langages de mots infinis. Nous présentons les automates de Büchi munis d'un nouveau critère de reconnaissance permettant de reconnaître des mots de longueur infinie.
Un *automate fini* $\mathcal{A}$ est la donnée de $(\Sigma, Q, \delta, I, F)$ où Σ désigne un alphabet, Q un ensemble fini d'états, $I \subseteq Q$ un ensemble d'états initiaux, $F \subseteq Q$ un ensemble d'états de reconnaissance et $\delta \subseteq (Q \times \Sigma \times Q)$ un ensemble de transitions.

Un automate fini $\mathcal{A}$ est *déterministe* si son ensemble I d'états initiaux est singleton et si δ est une fonction de $Q \times \Sigma$ dans Q.

Pour tout ω-mot $w \in \Sigma^\omega$, une *lecture* λ de w dans $\mathcal{A}$ est une suite infinie d'états $(\lambda_i)_{i>0}$ telle que : $\forall i > 0,\ (\lambda_i, w_i, \lambda_{i+1}) \in \delta$. On note $Inf(\lambda)$ l'ensemble des états qui apparaissent infiniment souvent dans la lecture λ de w :

$$Inf(\lambda) = \{q,\ card(\lambda^{-1}(q)) \text{ est infini}\}$$

Comme Q est fini, $\mathit{Inf}(\lambda)$ n'est jamais vide.
Un mot $w \in \Sigma^\omega$ est *reconnu* par $\mathcal{A}$ s'il existe une lecture *réussie* λ de w, c'est-à-dire telle que $\lambda_1 \in I$ et $\mathit{Inf}(\lambda) \cap F \neq \emptyset$. L'automate $\mathcal{A}$ muni de ce critère de reconnaissance est un *automate de Büchi*.
L'*ω-langage reconnu* par $\mathcal{A}$ est l'ensemble des ω-mots reconnus par $\mathcal{A}$. On le note $L_\omega(\mathcal{A})$.

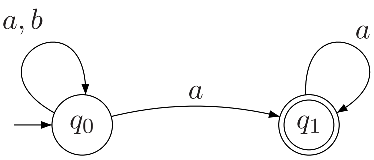

FIGURE 3.5 – L'automate fini $\mathcal{A}$

Exemple Soit $\mathcal{A}$ l'automate $(\Sigma, Q, \delta, I, F)$ tel que $Q = \{q_0, q_1\}$, $I = \{q_0\}$, $F = \{q_1\}$ et δ comme explicité sur la Figure 3.5. Le langage reconnu par $\mathcal{A}$ est $L(\mathcal{A}) = (a+b)^*a$ et l'ω-langage reconnu par $\mathcal{A}$ est $L_\omega(\mathcal{A}) = (a+b)^*a^\omega$.

Exercice 26 Donnez les automates de Büchi reconnaissant les ω-langages suivants : $(b^*a)^\omega$, $(ba^*)^\omega$ et $\{w \in \Sigma^\omega,\ |w|_a = |w|_b$ est infini$\}$. Repérez parmi eux les automates déterministes.

Dorénavant, on dira simplement *automates* pour les automates finis classiques, i.e. munis du critère de reconnaissance pour les langages (finitaires), et *automates de Büchi* pour les automates finis qui reconnaissent des ω-langages (infinitaires donc).

3.3.1 Reconnaissabilité

Nous allons établir ici le lien entre les langages reconnaissables par automate de Büchi et les langages rationnels de mots infinis définis en Section 3.2.2.

Définition $\mathit{Rec}(\Sigma^\omega)$ est la classe des ω-langages reconnaissables *i.e.* l'ensemble des ω-langages reconnus par automate de Büchi sur l'alphabet Σ.

Exercice 27 Montrez que toute partie reconnaissable non-vide de Σ^ω contient un mot ultimement périodique.

La Proposition 18 à venir énonce l'égalité entre la classe des ω-langages rationnels et celles des ω-langages reconnaissables par automate de Büchi. On commence par donner deux lemmes afin de prouver le premier sens de l'inclusion. Rappelons qu'un ω-langage rationnel L est de la forme $\bigcup_{i=1}^{n} A_i B_i^{\omega}$ avec $A_i, B_i \in Rat(\Sigma^*)$.

Lemme 8 *Pour tout $A, B \in Rat(\Sigma^*)$, l'ω-langage $AB^{\omega} \in Rec(\Sigma^{\omega})$.*

Démonstration Sans perte de généralité, on pose $\varepsilon \notin B$. On suppose aussi que $\varepsilon \notin A$ (sinon comme $AB^{\omega} = (AB)B^{\omega}$, on prend AB à la place de A comme ça, $\varepsilon \notin AB$).
Il existe un automate fini reconnaissant A ayant un seul état de reconnaissance q_f d'où aucune flèche ne sort (facile à démontrer). L'automate est alors non-déterministe.
Il existe un automate fini reconnaissant B^* ayant un seul état de reconnaissance qui est aussi l'unique état initial p_0. On rassemble les deux automates en confondant q_f et p_0. □

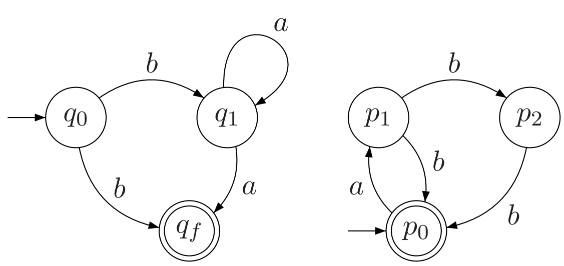

FIGURE 3.6 – Construction d'un automate de Büchi pour AB^{ω}

Exemple Visualisez une telle construction sur la Figure 3.6 pour les langages rationnels $A = ba^*$ et $B = ab + abb$.

Lemme 9 *La classe* $Rec(\Sigma^\omega)$ *est fermée par union.*

Démonstration Soit $\mathcal{A} = (\Sigma, Q, \delta, I, F)$ et $\mathcal{A}' = (\Sigma, Q', \delta', I', F')$ deux automates de Büchi. On crée l'automate-produit $\mathcal{A}'' = (\Sigma, Q \times Q, \delta'', I \times I', (F \times Q') \cup (Q \times F'))$ avec $\delta'' = \{((q, q'), \sigma, (p, p'))/\ \sigma \in \Sigma, (q, \sigma, p) \in \delta$ et $(q', \sigma, p') \in \delta'\}$. On a bien $L_\omega(\mathcal{A}'') = L_\omega(\mathcal{A}) \cup L_\omega(\mathcal{A}')$. $\square$

De plus, si $\mathcal{A}$ et $\mathcal{A}'$ sont déterministes, l'automate obtenu $\mathcal{A}''$ est également déterministe. A noter qu'un produit d'automates analogue ne fournit pas d'automate pour l'intersection. Toutefois, la classe $Rec(\Sigma^\omega)$ est fermée par intersection (cf. Proposition 7).

Exercice 28 Essayons cette construction pour l'union : $L = (a^*b)^\omega \cup (b^*a)^\omega$. Qu'advient-il si on tente la construction pour l'intersection au lieu de l'union ?

Voici l'analogue du Théorème de Kleene pour les ω-langages :

Proposition 18 *L'ensemble des ω-langages rationnels coïncide avec l'ensemble des ω-langages reconnaissables. Autrement dit, sur l'alphabet Σ :*

$$Rat(\Sigma^\omega) = Rec(\Sigma^\omega)$$

Démonstration Sachant qu'un élément L de $Rat(\Sigma^\omega)$ est de la forme $\bigcup_{i=1}^{n} A_i B_i^\omega$ avec $A_i, B_i \in Rat(\Sigma^*)$, l'inclusion $Rat(\Sigma^\omega) \subseteq Rec(\Sigma^\omega)$ est obtenue grâce aux Lemmes 8 et 9.
Réciproquement, prenons un automate de Büchi $\mathcal{A} = (\Sigma, Q, \delta, I, F)$, on note $L = L_\omega(\mathcal{A})$. Pour chaque couple d'états (q, q') de Q^2, on crée les automates $\mathcal{A}_{q,q'}$ ainsi : $\mathcal{A}_{q,q'} = (\Sigma, Q, \delta, q, \{q'\})$. Un mot w appartient à L si et seulement si il existe deux états $q \in I$ et $q' \in F$ de sorte que la lecture de w débute dans q et passe infiniment souvent par l'état q'. Ainsi, $L_\omega(\mathcal{A}) = \bigcup_{q \in I, q' \in F} L(\mathcal{A}_{q,q'}).\left(L(\mathcal{A}_{q',q'})\right)^\omega$. $\square$

D'après ce qui précède, il est clair qu'un mot w de $L_\omega(\mathcal{A})$ admet une infinité de préfixes dans $L(\mathcal{A})$:

Corollaire 2 *Pour tout automate fini $\mathcal{A}$ muni tour à tour des critères de reconnaissance des mots finis (automate classique) et infinis (automate de Büchi), on a :*

$$L_\omega(\mathcal{A}) \subseteq \overrightarrow{L(\mathcal{A})}$$

3.3.2 ω-langages déterministes

Tous les ω-langages de $Rec(\Sigma^\omega)$ ne sont pas reconnaissables par automate de Büchi déterministe. On appelle $Drat(\Sigma^\omega)$ la classe des ω-langages rationnels reconnus par un automate de Büchi déterministe et on cherche à les caractériser dans $Rat(\Sigma^\omega)$. Les éléments de $Drat(\Sigma^\omega)$ sont appelés les *ω-langages déterministes.*

Exemple L'automate fini $\mathcal{A}$ de la Figure 3.5 n'était pas déterministe. En revanche, l'automate fini $\mathcal{B}$ de la Figure 3.7 est déterministe. $\mathcal{B}$ est la version déterministe de $\mathcal{A}$ dans la mesure où $L(\mathcal{B}) = L(\mathcal{A}) = (a+b)^*a$. Toutefois, l'$\omega$-langage $L_\omega(\mathcal{B}) = (b^*a)^\omega$ reconnu par $\mathcal{B}$ est différent de $L_\omega(\mathcal{A}) = (a+b)^*a^\omega$.

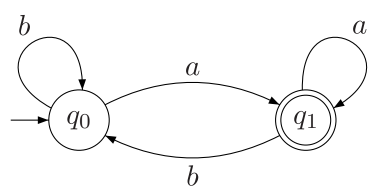

FIGURE 3.7 – L'automate fini déterministe $\mathcal{B}$

Proposition 19 *Si $\mathcal{A}$ est un automate de Büchi déterministe alors $L_\omega(\mathcal{A}) = \overrightarrow{L(\mathcal{A})}$*

Démonstration D'après le Corollaire 2, on sait que $L_\omega(\mathcal{A}) \subseteq \overrightarrow{L(\mathcal{A})}$. Réciproquement, soit $w \in \overrightarrow{L(\mathcal{A})}$. Il existe une infinité d'entiers n distincts tels que les préfixes de w de longueur n notés $w_n \in L(\mathcal{A})$. Comme l'automate est déterministe, il existe au moins un état $q \in F$ tel qu'une infinité de n vérifie $\delta(q_0, w_n) = q$. □

Dans la démonstration précédente, si l'automate $\mathcal{A}$ n'était pas déterministe, il pourrait y avoir $n_2 > n_1$ tel que $\delta(q_0, w_{n_1}) = q$ et $\delta(q_0, w_{n_2}) = q$ sans que la lecture de w_{n_2} ne passe par q.
Voici une conséquence immédiate de la Proposition 19 :

Corollaire 3 *Soit L un ω-langage de Σ^ω. L appartient à $Drat(\Sigma^*)$ si et seulement si il existe un langage rationnel K de Σ^* tel que :*

$$L = \overrightarrow{K}$$

Exemple Soit $\mathcal{B}$ l'automate de la Figure 3.7. Il est déterministe donc nous avons la relation suivante entre $L(\mathcal{B}) = (a+b)^*a$ et $L_\omega(\mathcal{B}) = (b^*a)^\omega$:

$$(b^*a)^\omega = \overrightarrow{(a+b)^*a}$$

Exercice 29 Chacun des ω-langages suivants : $(b^*a)^\omega$, $(ba^*)^\omega$ et $\{w \in \Sigma^\omega,\ |w|_a = |w|_b$ est infini$\}$ est reconnu par un automate de Büchi déterministe comme vu précédemment. Trouvez les langages dont ces ω-langages sont les limites.

Proposition 20 *La classe $Drat(\Sigma^\omega)$ est fermée par union.*

Démonstration Analogue à celle du Lemme 9 à partir du moment où $Rat(\Sigma^\omega) = Rec(\Sigma^\omega)$ est assurée par la Proposition 18. Le déterminisme est bien conservé. □

Proposition 21 *La classe $Drat(\Sigma^\omega)$ est fermée par intersection.*

Démonstration Soit $\mathcal{A} = (\Sigma, Q, \delta, q_0, F)$ et $\mathcal{A}' = (\Sigma, Q', \delta', q'_0, F')$ deux automates de Büchi déterministes. On a $L_\omega(\mathcal{A}) = \bigcup_{q\in F} L_\omega(\mathcal{A}_{q_0,q})$ et $L_\omega(\mathcal{A}') = \bigcup_{q'\in F'} L_\omega(\mathcal{A}'_{q'_0,q'})$ avec $\mathcal{A}_{x,y}$ désignant l'automate $(\Sigma, Q, \delta, x, \{y\})$.
Ainsi, $L_\omega(\mathcal{A}) \cap L_\omega(\mathcal{A}') = \bigcup_{q\in F, q'\in F'} \Big(L_\omega(\mathcal{A}_{q_0,q}) \cap L_\omega(\mathcal{A}'_{q'_0,q'}) \Big)$. Il suffit de montrer que la grande parenthèse est dans $Drat(\Sigma^\omega)$. Sans perte de généralité, on suppose que $\mathcal{A}$ et $\mathcal{A}'$ ont chacun un seul état de reconnaissance respectivement t et t'. Dans la lecture d'un mot accepté par $\mathcal{A}$, l'idée est que l'état t sera considéré comme un état de reconnaissance pour l'intersection que si on a traversé t' depuis la dernière fois où on a visité t.
On construit l'automate $\mathcal{A}'' = (\Sigma, Q'', \delta'', (q_0, q'_0), \{t\}\times Q')$ avec $Q'' = ((Q \cup \overline{Q})\times Q')$ pour reconnaître l'intersection. Un élément $\overline{t}$ signifie qu'on est passé de t à t sans passer par t'. Reste à expliciter les transitions de δ'' :

$$\delta''((q, q'), \sigma) = \begin{cases} (\overline{\delta(q,\sigma)}, \delta'(q',\sigma)) & \text{si } q = t \text{ et } q' \neq t' \\ (\delta(q,\sigma), \delta'(q',\sigma)) & \text{sinon} \end{cases}$$

$$\delta''((\overline{q}, q'), \sigma) = \begin{cases} (\overline{\delta(q,\sigma)}, \delta'(q',\sigma)) & \text{si } q' \neq t' \\ (\delta(q,\sigma), \delta'(q',\sigma)) & \text{sinon} \end{cases}$$

On a bien $L_\omega(\mathcal{A}'') = L_\omega(\mathcal{A}) \cap L_\omega(\mathcal{A}')$. □

Dans le cas où $\mathcal{A}$ et $\mathcal{A}'$ ne sont pas déterministes, nous obtenons tout de même une démonstration du fait que $Rat(\Sigma^\omega)$ est close par intersection.

Exercice 30 Montrez que le langage $L = (a + b)^*a^\omega$ de la Figure 3.5 n'est pas un langage déterministe car il ne peut pas être mis sous forme de la limite d'un langage rationnel (ni même d'un langage non-rationnel en fait).

Corollaire 4 *Soit Σ un alphabet donné,*

$$Drat(\Sigma^\omega) \subsetneq Rat(\Sigma^\omega)$$
$Drat(\Sigma^\omega)$ n'est pas fermée par complémentation.

En effet, l'ω-langage rationnel $L = (a+b)^*a^\omega$ n'est pas déterministe alors que son complémentaire $\overline{L} = (a^*b)^\omega$ est dans $Drat(\Sigma^\omega)$.

La classe des ω-langages rationnels déterministes peut être définie au moyen de la classe des langages rationnels :

Proposition 22 *[5]*
Soit L un ω-langage. Les propriétés suivantes sont équivalentes :
- *(i) $L \in Drat(\Sigma^\omega)$*
- *(ii) $L = \overrightarrow{M}$ où $M \in Rat(\Sigma^*)$*
- *(iii) $L = \bigcup_{i=1}^{n} A_i B_i^\omega$, où $n \geq 0$ et $\forall i, 1 \leq i \leq n,\ A_i, B_i$ sont des langages préfixes rationnels.*

Démonstration (iii)⇒(ii) $L = \bigcup_{i=1}^{n} A_i B_i^\omega$ avec A_i, B_i préfixes et rationnels.
Soit $M = \bigcup_{i=1}^{n} A_i B_i^*$
$\overrightarrow{M} = \bigcup_{i=1}^{n} \overrightarrow{A_i B_i^*}$
$= \bigcup_{i=1}^{n} A_i B_i^\omega$
donc $\overrightarrow{M} = L$.
(ii)⇒(i) $L = \overrightarrow{M}$ avec $M \in Rat(\Sigma^*)$. Soit $\mathcal{A}$ un automate fini déterministe tel que $L(\mathcal{A}) = M$. D'après la Proposition 19, $L_\omega(\mathcal{A}) = \overrightarrow{L(\mathcal{A})} = \overrightarrow{M} = L$.
(i)⇒(iii) Soit $\mathcal{A} = (\Sigma, Q, \delta, q_0, F)$ un automate déterministe tel que $L_\omega(\mathcal{A}) = L$. Pour tout $q \in F$, on crée les automates $\mathcal{A}_{q_0,q} = (\Sigma, Q, \delta, q_0, \{q\})$ et $\mathcal{A}_{q,q} = (\Sigma, Q, \delta, q, \{q\})$.
On note : $A_q = L(\mathcal{A}_{q_0,q}) \in Rat(\Sigma^*)$
$B_q = L(\mathcal{A}_{q,q}) \in Rat(\Sigma^*)$
$A'_q = A_q \setminus A_q\Sigma^+ \in Rat(\Sigma^*)$
$B'_q = B_q \setminus B_q\Sigma^+ \in Rat(\Sigma^*)$
Par construction, A'_q et B'_q sont préfixes. Reste à montrer que : $L_\omega(\mathcal{A}) = \bigcup_{q\in F} A'_q(B'_q)^\omega$.
D'une part, $A'_q(B'_q)^\omega \subseteq A_q(B_q)^\omega \subseteq L_\omega(\mathcal{A})$. Réciproquement, $L_\omega(\mathcal{A}) \subseteq \bigcup_{q\in F} A'_q(B'_q)^\omega$. En effet, soit $w \in L_\omega(\mathcal{A})$, il existe $q \in F$ tel qu'une infinité de i vérifient $\delta^*(q_0, w_i) = q$. L'ω-mot w se décompose en $w = \alpha\beta_1\beta_2...$ avec $\alpha \in A'_q$ et $\forall i \geq 0,\ \beta_i \in B'_q$ donc $w \in A'_q(B'_q)^\omega$. □

Si $\mathcal{A}$ est un automate de Büchi non-déterministe tel que $L_\omega(\mathcal{A}) = \overrightarrow{L(\mathcal{A})}$, alors $L_\omega(\mathcal{A}) \in Drat(\Sigma^*)$. Mais si $L_\omega(\mathcal{A}) \neq \overrightarrow{L(\mathcal{A})}$, on ne peut pas en déduire que $L_\omega(\mathcal{A}) \notin Drat(\Sigma^*)$.

Exemple Soit $\mathcal{A}$ l'automate du haut de la Figure 3.8. $L(\mathcal{A}) = \varepsilon + b\Sigma^*$ et $L_\omega(\mathcal{A}) = (b\Sigma^*)^\omega$. On a $\overrightarrow{L(\mathcal{A})} = b\Sigma^\omega \neq (b\Sigma^*)^\omega = L_\omega(\mathcal{A})$ mais il existe un automate de Büchi déterministe $\mathcal{B}$, en bas de la même figure, qui reconnaît $L_\omega(\mathcal{A})$ donc $L_\omega(\mathcal{A}) \in Drat(\Sigma^\omega)$.

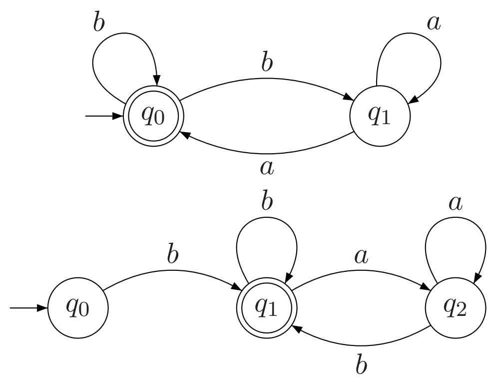

FIGURE 3.8 – Automates non-déterministe $\mathcal{A}$ et déterministe $\mathcal{B}$ tels que $L_\omega(\mathcal{A}) = L_\omega(\mathcal{B})$

Théorème 13 *(Büchi-Mac Naughton)*

$$Rat(\Sigma^\omega) = (Drat(\Sigma^\omega))^B$$

où ^{B} *est mis pour la clôture booléenne,* [1].

Ce théorème est en fait issu des deux théorèmes suivants :

Théorème 14 *(Büchi) [3]*
La classe $Rat(\Sigma^\omega)$ *est une algèbre de Boole.*

Il suffirait de montrer que $Rat(\Sigma^\omega)$ est close par complémentation puisque par définition, la classe $Rat(\Sigma^\omega)$ est close par union.

Théorème 15 *(Mac Naughton) [13]*

$$Rat(\Sigma^\omega) \subseteq (Drat(\Sigma^\omega))^B$$

Pour démontrer un tel résultat, nous devons introduire les automates de Muller ...

1. Une algèbre de Boole est un ensemble clos par union finie, intersection finie et complémentaire.

3.4 Automates de Muller

L'enjeu est d'étendre le modèle des automates de Büchi afin de parvenir à une caractérisation de la classe des ω-langages rationnels en termes d'automates déterministes.

Pour ce faire, on introduit un autre critère de reconnaissance des ω-mots. L'ensemble de reconnaissance T de l'automate n'est plus composé d'états mais d'un sous-ensemble de l'ensemble des états. Un ω-mot est reconnu par un automate de Muller si l'ensemble des états apparaissant infiniment souvent dans la lecture de w est précisément un élément de T.

Plus formellement, un *automate de Muller* $\mathcal{A}$ est la donnée de $(\Sigma, Q, \delta, q_0, T)$ où Σ désigne un alphabet, Q un ensemble fini d'états, $\delta \subseteq (Q \times \Sigma \times Q)$ la relation de transition, $q_0 \in Q$ l'état initial et $T \subseteq 2^Q$ est l'ensemble des états de reconnaissance, [2].

Une *lecture* d'un ω-mot w de Σ^ω est définie de même que pour les automates de Büchi (*cf.* Section 3.3). Un ω-mot w est *reconnu* par $\mathcal{A}$ s'il existe une lecture *réussie* λ de w, c'est-à-dire telle que $\lambda_0 = q_0$ et :

$$\exists P \in T,\ \mathit{Inf}(\lambda) = P$$

L'*ω-langage reconnu* par $\mathcal{A}$ est l'ensemble des ω-mots reconnus par $\mathcal{A}$. On le note $L_\omega(\mathcal{A})$.

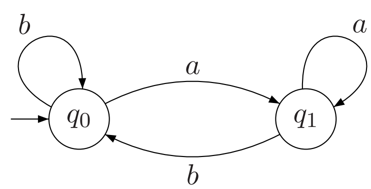

FIGURE 3.9 – Automate de Muller déterministe $\mathcal{A}$ avec $T = \{\{q_1\}\}$

Exemple L'automate de Muller $\mathcal{A} = (\Sigma, Q, \delta, q_0, T)$ avec $\Sigma = \{a, b\}$, $Q = \{q_0, q_1\}$, $T = \{\{q_1\}\}$ avec δ explicite sur la Figure 3.9 est déterministe. Il reconnaît le langage $L = (a + b)^* a^\omega$ qui n'est pas reconnu par un automate de Büchi déterministe.

2. Partant, on ne peut plus faire figurer les ensembles de reconnaissance constituant T sur le schéma d'un automate de Muller. Il faut donner explicitement cet ensemble T, appelé sa *table*.

Bibliographie

[1] J. Berstel and D. Perrin. *Theory of codes.* Academic Press, 1985.

[2] J. Berstel, D. Perrin, and Ch. Reutenauer. *Codes and automata*, volume 129 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, New-York, 2010. 619 pages.

[3] J.R. Büchi. On a decision method in restricted second order arithmetics. In *International Congress on Logic, Methodology and Philosophy of Science*, pages 1–11. Stanford University Press, 1960.

[4] M. Crochemore, C. Hancart, and T. Lecroq. *Algorithms on strings.* Cambridge University Press, 2007.

[5] S. Eilenberg. *Automata, Languages and Machines*, volume A. Academic Press, 1974.

[6] C. C. Elgot. Decision problems of finite automata design and related arithmetics. *Trans. of Amer. Math. Soc.*, 98 :21–51, 1961.

[7] J.E. Hopcroft and J.D. Ullman. *Decidability*, chapter 14. Addison-Wesley, 1979.

[8] S. Julia, A. Malapert, and J. Provillard. A synergic approach to the minimal uncompletable words problem. *Journal of Automata, Languages and Combinatorics*, 22(4) :271–286, 2017.

[9] S. C. Kleene. Representation of events in nerve nets and finite automata. In *Automata studies*, pages 3–40. Princeton, 1956.

[10] M. Lothaire. *Combinatorics on words.* Reading, Massachusetts : Addison-Wesley, 1983.

[11] M. Lothaire. *Algebraic Combinatorics on Words*, volume 90 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, 2002. 504 pages.

[12] M. Lothaire. *Applied Combinatorics on Words.* Cambridge University Press, 2005.

[13] R. Mac Naughton. Testing and generating infinite sequences by a finite automaton. *Information and Control*, 9 :521–530, 1966.

[14] D. Perrin and J.-É. Pin. *Infinite words*. Elsevier, 2004.

[15] A. Restivo. Some remarks on complete subsets of a free monoid. *Quaderni de La Ricerca Scientifica*, 109 :19–25, 1981.

[16] J. Shallit. *A Second Course in Formal Languages and Automata Theory*. Cambridge University Press, 2009.

[17] L. Staiger. *Handbook of Formal Languages*, chapter ω-languages, pages 339–387. Springer, 1997.

[18] H. Straubing. *Finite Automata, Formal Logic and Circuit Complexity*. Birkhäuser, 1994.

[19] W. Thomas. Automata on infinite objects. In *Handbook of Theoretical Computer Science*, volume B, chapter 4. Elsevier Science Publishers, 1990.

[20] W. Thomas. *Handbook of Formal Languages*, chapter Languages, automata and logic, pages 389–455. Springer, 1997.