Langages formels (11X003) - Automne 2024

5. Lemme de l'étoile

Enseignant: Arnaud Casteigts

Assistants: A.-Q. Berger & M. De Francesco
Monitrices: L. Heiniger & A. Tekkoyun

Aussi appelé lemme de gonflement ou lemme de pompage, le **lemme de l'étoile** est un outil central de la théorie des langages formels. Il désigne une propriété que *tous* les langages réguliers *doivent* satisfaire. Par conséquent, si l'on montre qu'un langage ne satisfait pas cette propriété, cela montre qu'il n'est pas régulier.

5.1 Lemme de l'étoile

Le lemme de l'étoile établit que tout langage régulier doit satisfaire une certaine propriété. Quelle est cette propriété ?

Pour comprendre, le plus simple est de réfléchir en termes d'automates. Si un langage est régulier, alors il existe un automate fini qui le reconnaît (et même un AFD). Soit L un langage régulier et $A(L)$ un AFD qui reconnaît L. Si on regarde le chemin pris par cet automate pour reconnaître un mot $w \in L$, il y a deux possibilités :

- Soit $A(L)$ accepte w sans jamais passer deux fois par le même état,
- Soit $A(L)$ accepte w en repassant au moins une fois par le même état.

Si la longueur du mot est grande, par exemple plus grande que le nombre k d'états de l'automate, alors nécessairement, on est dans le deuxième cas et il y aura au moins un état répété pendant la lecture de w, disons l'état q. On peut alors découper le mot w en trois morceaux (facteurs) $x \cdot y \cdot z$ tels que :

- x est un préfixe lu avant d'arriver sur l'état q,
- y est un facteur dont la lecture commence et termine sur l'état q,
- z est un suffixe dont la lecture commence sur l'état q et termine sur un état final.

Il est possible, dans des cas particuliers, que x ou z soient des mots vides, par exemple si q est un état initial ou final (respectivement). Par contre, $|y|$ est strictement positif car nous avons supposé que l'exécution passe *plusieurs* fois par l'état q. Que se passerait-il si au lieu de lire $x \cdot y \cdot z$, on lisait le mot $x \cdot y \cdot y \cdot z$? Le fait que la lecture de y commence et termine sur le même état implique que ce mot sera forcément accepté aussi (avec une

répétition supplémentaire sur l'état q). Il en va de même pour $x \cdot y \cdot y \cdot y \cdot z$, et en fait, tous les mots de la forme $x \cdot y^i \cdot z$ (et même $x \cdot y^0 \cdot z = xz$, d'ailleurs).

Récapitulons :

Lemme 5.1 (Lemme de l'étoile)**.** *Si un langage L est régulier, alors il existe une longueur k au delà de laquelle tout mot $w \in L$ peut être écrit $x \cdot y \cdot z$ avec :*

1. $|y| > 0$,
2. $x \cdot y^i \cdot z \in L$ *pour tout* $i \in \mathbb{N}$.
3. $|x \cdot y| \leq k$

Nous avons déjà expliqué les deux premières propriétés. La troisième peut sembler plus artificielle, mais elle s'avère souvent utile et on peut toujours la satisfaire en choisissant x et y de manière appropriée.

5.2 Utilisation

Ainsi, tout langage régulier doit satisfaire le lemme de l'étoile. On peut donc l'utiliser pour montrer qu'un langage L n'est pas régulier. La démarche à suivre est classique : par l'absurde, on suppose d'abord que L *est* régulier, ce qui implique que le lemme de l'étoile est satisfait, puis on utilise le lemme de l'étoile pour montrer une contradiction : l'automate reconnaissant L acceptera des mots qui ne sont pas dans L. Le langage L n'est donc pas régulier.

5.2.1 $L = \{\mathtt{a}^n\mathtt{b}^n \mid n \in \mathbb{N}\}$ sur l'alphabet $\Sigma = \{a, b\}$

Prenons l'exemple du langage $L = \{\mathtt{a}^n\mathtt{b}^n \mid n \in \mathbb{N}\}$ sur l'alphabet $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, autrement dit le langage $L = \{\varepsilon, \mathtt{ab}, \mathtt{aabb}, \mathtt{aaabbb}, \mathtt{aaaabbbb}, \dots\}$, qui est infini.

Si L est régulier, alors le lemme de l'étoile nous dit qu'il existe une longueur k au delà de laquelle tout mot du langage peut être décomposé sous la forme $x \cdot y \cdot z$, avec $|y| > 0$ et $x \cdot y^i \cdot z \in L$ pour tout i. L étant infini, il existe forcément des mots ayant au moins cette longueur là. Prenons-en un, disons w avec $|w| > k$, et décomposons sous forme $x \cdot y \cdot z$. Il se peut qu'il y ait plusieurs décompositions possibles. Mais quelle que soit la décomposition, il y aura trois possibilités :

1. y n'a que des $\mathtt{a}$,
2. y n'a que des $\mathtt{b}$,
3. y commence par des $\mathtt{a}$ et termine par des $\mathtt{b}$.

Examinons chaque cas. Si y n'a que des $\mathtt{a}$, alors $x \cdot y^2 \cdot z$ aura plus de $\mathtt{a}$ que de $\mathtt{b}$, ce qui contredit la propriété 2 du lemme de l'étoile. Si y n'a que des $\mathtt{b}$, alors $x \cdot y^2 \cdot z$ aura plus

de `b` que de `a` (idem). Enfin, si y commence par des `a` et termine par des `b`, alors $x \cdot y^2 \cdot z$ alternera des `a`, puis des `b`, puis des `a`, puis des `b`, il n'est donc pas non plus dans L (là encore, contradiction). Dans chacun des cas, on arrive à la contradiction que L contient un mot... qui n'est pas dans L (glups). Le langage L ne peut donc pas être régulier.

5.2.2 $L' = \{w \mid w$ a autant de `a` que de `b`$\}$ sur l'alphabet $\Sigma = \{a, b\}$

En exercices, vous montrez que L' n'est pas régulier, en utilisant le lemme de l'étoile et en utilisant cette fois la troisième propriété ($|xy| \leq k$). En attendant, on peut démontrer cela autrement, en faisant un lien utile avec le langage L de la section 5.2.1.

Définissons d'abord un autre L'' correspondant à l'expression régulière a^*b^*. Ce langage est bien régulier, puisqu'on l'a définit à partir d'une expression régulière.

Que vaut le langage $L' \cap L''$? Autrement dit tous les mots qui ont autant de `a` que de `b` (mots de L') et (intersection) qui ont d'abord des a puis des b (mots de L''). Réponse : les mots de cette intersection sont exactement les mots de L, on a donc $L' \cap L'' = L$.

Il se trouve que l'intersection de deux langages réguliers est toujours un langage régulier. Sachant que L'' est régulier, on peut donc dire que si L' était régulier, alors $L' \cap L'' = L$ le serait aussi. Mais on sait déjà que L n'est pas régulier, donc L' ne l'est pas non plus.

5.3 Conclusion

Certain langages ne sont pas réguliers et ne peuvent donc pas être reconnus par des automates finis. Par exemple, le langage $\{a^n b^n \mid n \in \mathbb{N}\}$ n'est pas régulier. Le langage de tous les palindromes est un autre exemple. Plus tard dans le cours, nous étudierons des modèles de machines plus puissantes, qui permettent de reconnaître ces langages. Puis nous verrons (à leur tour) les limites de ces machines. Ultimement, nous montrerons que même les ordinateurs d'aujourd'hui (et ceux de demain...) ne peuvent pas reconnaître tous les langages.