Langages formels (11X003) - Automne 2023

5. Lemme de l'étoile

Enseignant: Arnaud Casteigts *Assistant: Alexandre-Quentin Berger*

Remarque : la fin du cours sur les expressions régulières a été ajoutée aux notes de cours de la semaine dernière (Cours 4 - Expressions régulières, Section 4.4). Les notes ci-dessous se concentrent sur l'outil du jour : le lemme de l'étoile.

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 permet de montrer qu'un langage donné n'est pas régulier et qu'il ne peut donc pas être reconnu par un automate fini (déterministe ou non), ni représenté par une expression régulière.

5.1 Lemme de l'étoile

Le lemme de l'étoile établit que tout langage régulier doit satisfaire une certaine propriété. Si un langage n'a pas cette propriété, il ne peut donc pas être régulier. 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 après être passé plusieurs 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. Par contre, $|y|$ est strictement positif car nous avons choisi k

pour forcer l'automate à passer 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 $x \cdot y^0 \cdot z = xz$, au passage).

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, puis on arrive à une contradiction du lemme de l'étoile. Ce langage ne peut donc pas être régulier.

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, et décomposons sous forme $x \cdot y \cdot z$. Quelle que soit la manière de le décomposer, la définition de L implique 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 $\mathtt{b}$ que de $\mathtt{a}$ (idem). Enfin, si y commence par des $\mathtt{a}$ et termine par des $\mathtt{b}$, alors $x \cdot y^2 \cdot z$ alternera des $\mathtt{a}$, puis des $\mathtt{b}$, puis des $\mathtt{a}$, puis des $\mathtt{b}$, il n'est donc pas non plus dans L (là encore, contradiction). Dans chacun des cas, le lemme de l'étoile est contredit, L ne peut donc pas être régulier.

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 (ou de demain...) ne peuvent pas reconnaître certain langages.